U.S. DEPARTMENT OF COMMERCE
National Technical Information Service
PB80-152838

Performance Measures and Standards for Air Quality Simulation Models

Systems Applications, Inc, San Rafael, CA

Prepared for

Environmenal Protection Agency, Research Triangle Park, NC

Oct 79

United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/4-79-032
October 1979

Air

PB80-152338

EPA

Performance Measures and Standards for Air Quality Simulation Models

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-450/4-79-032	2.	3. RECIPIENT'S ACCESSION NO. PB80-152838
4. TITLE AND SUBTITLE Performance Measures and Standards for Air Quality Simulation Models		5. REPORT DATE October 1979
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) S. R. Hayes		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Systems Applications, Incorporated 950 Northgate Drive San Rafael, California 94903		10. PROGRAM ELEMENT NO.
		11. CONTRACT/GRANT NO. 68-02-2593
12. SPONSORING AGENCY NAME AND ADDRESS Office of Air Quality Planning and Standards U. S. Environmental Protection Agency Research Triangle Park, North Carolina 27711		13. TYPE OF REPORT AND PERIOD COVERED Final Report
		14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Currently there are no standardized guidelines for evaluating the performance of air quality simulation models. In this report we develop a conceptual framework for objectively evaluating model performance. We define five attributes of a well-behaving model: accuracy of the peak prediction, absence of systematic bias, lack of gross error, temporal correlation, and spatial alignment. The relative importance of these attributes is shown to depend on the issue being addressed and the pollutant being considered. Acceptability of model behavior is determined by calculating several performance "measures" and comparing their values with specific "standards." Failure to demonstrate a particular attribute may or may not cause a model to be rejected, depending on the issue and pollutant.

Comprehensive background material is presented on the elements of the performance evaluation problem: the types of issues to be addressed, the classes of models to be used along with the applications for which they are suited, and the categories of performance measures available for consideration. Also, specific rationales are developed on which performance standards could be based. Guidance on the interpretation of performance measure values is provided by means of an example using a large, grid-based air quality model.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollution Turbulent Diffusion Mathematical Models Computer Models Atmospheric Models	Dispersion Air Quality Simulation Model Model Validation Model Evaluation	

18. DISTRIBUTION STATEMENT Release Unlimited	19. SECURITY CLASS *(This Report)* None	21. NO. OF PAGES
	20. SECURITY CLASS *(This page)* None	22. PRICE

PA Form 2220-1 (9-73)

EPA-450/4-79-032

Performance Measures and Standards for Air Quality Simulation Models

by

S.R. Hayes

Systems Applications, Incorporated
950 Northgate Drive
San Rafael, California 94903

Contract No. 68-02-2593

EPA Project Officer: Russell F. Lee

Prepared for

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Air, Noise, and Radiation
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711

October 1979

i-a

DISCLAIMER

This report has been reviewed by the Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Enviornmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

is developed

are defined

ABSTRACT

Currently there are no standardized guidelines for evaluating the performance of air quality simulation models. In this report we develop a conceptual framework for objectively evaluating model performance. We define five attributes of a well-behaving model: accuracy of the peak prediction, absence of systematic bias, lack of gross error, temporal correlation, and spatial alignment. The relative importance of these attributes is shown to depend on the issue being addressed and the pollutant being considered. Acceptability of model behavior is determined by calculating several performance "measures" and comparing their values with specific "standards." Failure to demonstrate a particular attribute may or may not cause a model to be rejected, depending on the issue and pollutant.

Comprehensive background material is presented on the elements of the performance evaluation problem: the types of issues to be addressed, the classes of models to be used along with the applications for which they are suited, and the categories of performance measures available for consideration. Also, specific rationales are developed on which performance standards could be based. Guidance on the interpretation of performance measure values is provided by means of an example using a large, grid-based air quality model.

ACKNOWLEDGMENTS

A number of persons have generously provided their assistance and support to this project. Special thanks is due Philip Roth, whose foresight and leadership made this project possible. His perceptive advice and guidance contributed immeasurably to the results of this work.

Steven Reynolds and Martin Hillyer made many significant, insightful comments, which were greatly appreciated.

For their patience and diligence, grateful thanks is also due the members of the SAI support staff, particularly Marie Davis, Sue Bennett, Chris Smith, and Linda Hill.

CONTENTS

DISCLAIMER ii

ABSTRACT iii

ACKNOWLEDGMENTS. iv

LIST OF ILLUSTRATIONS vii

LIST OF TABLES xi

LIST OF EXHIBITS xiv

I INTRODUCTION I-1

A. Overview of the Problem I-2
B. Structure of the Report I-5

II SUMMARY II-1

A. Main Results II-1
B. Detailed Summary II-2
  1. Summary of Chapter III (Issues) II-2
  2. Summary of Chapter IV (Models) II-3
  3. Summary of Chapter V (Performance Measures) II-4
  4. Summary of Chapter VI (Performance Standards) II-14

III ISSUES REQUIRING MODEL APPLICATION III-1

A. A Perspective on the Issues III-1
  1. Federal Air Pollution Law III-2
  2. The Code of Federal Regulations III-3
B. Generic Issue Categories III-7
  1. The Issues: Their Classification III-8
  2. The Issues: Some Practical Examples and Their Implications for Air Pollution Modeling . . . III-10
  3. The Issues: A Prologue to the Next Chapter III-13

IV AIR QUALITY MODELS IV-1

A. Generic Model Categories IV-2

1. Rollback Category. IV-2
2. Isopleth Category. IV-4
3. Physico-Chemical Category. IV-5

B. Generic Issue/Model Combinations IV-16

C. Model/Application Combinations IV-22

D. Some Specific Air Quality Models IV-22

E. Air Quality Models: A Summary IV-25

V MODEL PERFORMANCE MEASURES V-1

A. The Comparison of Prediction with Observation. V-2

B. Generic Performance Measure Categories V-4

1. The Generic Measures V-5
2. Some Types of Variations Among Performance Measures V-10
3. Several Practical Considerations V-10

C. A Basic Distinction: Regional Versus Source-Specific Performance Measures V-15

D. Some Specific Performance Measures V-22

E. Matching Performance Measures to Issues and Models . . V-27

1. Performance Measures and Air Quality Issues. . . . V-27
2. Performance Measures and Air Quality Models. . . . V-33

F. Performance Measures: A Summary V-36

VI MODEL PERFORMANCE STANDARDS. VI-1

A. Performance Standards: A Conceptual Overview. VI-2

B. Performance Standards: Some Practical Considerations VI-4

1. Data Limitations VI-5
2. Time/Resource Constraints. VI-6
3. Variability of Analysis Requirements VI-6

C. Model Performance Attributes VI-7

D. Recommended Measures and Standards VI-12

1. Recommended Performance Measures VI-14
2. Recommended Performance Standards. VI-23
3. Summary Table of Recommended Measures and Standards. VI-30
4. Formulas for Calculating Performance Measures and Standards VI-32

VI MODEL PERFORMANCE STANDARDS (Continued)

E. A Sample Case: The SAI Denver Experience VI-39
  1. The Denver Modeling Problem VI-39
  2. Values of the Performance Measures. VI-40
  3. Interpreting the Performance Measure Values VI-45

F. Suggested Framework for a Draft Standard. VI-53

VII. RECOMMENDATIONS FOR FUTURE WORK VII-1

A. Areas for Technical Development VII-2
  1. Further Evaluation of Performance Measures VII-2
  2. Identification and Specification of Prototypical Point Source "Test Bed" Data Bases VII-2
  3. Examination of Performance Evaluation Procedure in Sparse-Data Point Source Applications VII-3
  4. Further Development of Rationales for Setting Performance Standards VII-4

B. Assessment of Institutional Implications VII-5

C. Documents To Be Compiled VII-6

APPENDICES

A IMPORTANT PARTS OF THE CODE OF FEDERAL REGULATIONS CONCERNING AIR PROGRAMS. A-1

B SOME SPECIFIC AIR QUALITY MODELS. B-1

C SOME SPECIFIC MODEL PERFORMANCE MEASURES C-1

D SEVERAL RATIONALES FOR SETTING MODEL PERFORMANCE STANDARDS D-1

REFERENCES R-1

ILLUSTRATIONS

II-1 Various Levels of Knowledge About Regional Concentrations . . . II-9

II-2 Various Levels of Knowledge About Specific-Source Concentrations . . . II-9

V-1 Various Levels of Knowledge About Regional Concentrations . . . V-6

V-2 Various Levels of Knowledge About Specific-Source Concentrations . . . V-7

V-3 Sample Regional Isopleth Diagram Illustrating Ozone Concentrations in Denver on 29 July 1975 for Hour 1200-1300 MST . . . V-17

V-4 Sample Specific-Source Isopleth Diagram Illustrating Concentrations Downwind of a Steady-State Gaussian Point Source . . . V-18

V-5 Concentration Isopleth Patterns for Various Source Types . . . V-20

V-6 Schematic of a Point Source Measurement Network . . . V-21

V-7 Locus of Possible Footprint Locations for an Elevated Point Source . . . V-21

VI-1 Orientation and Scaling of C_{AVE} and d* Axes on a Prediction-Observation Correlogram . . . VI-37

VI-2 Locations of Monitoring Stations in the Denver Metropolitan Region . . . VI-41

VI-3 Predicted and Observed Ozone Concentrations at Each Monitoring Station During the Day (Denver, 28 July 1976) . . . VI-42

VI-4 Correlogram of Ozone Observation-Prediction Pairs for Sample Case (Denver, 28 July 1976) . . . VI-46

VI-5 Normalized Deviations About the Perfect Correlation Line as a Function of Ozone Concentration (Denver, 28 July 1976) . . . VI-47

VI-6 Non-Normalized Ozone Deviations About the Perfect Correlation Line Compared with Instrument Errors (Data for 14 Hours and 8 Stations, Denver, 28 July 1976) . . . VI-48

VI-7 Non-Normalized Ozone Absolute Deviation About the Perfect Correlation Line Compared with Instrument Error (Data for 14 Hours and 8 Stations, Denver, 28 July 1976) . . . VI-49

VI-8 Ground-Traces of the Predicted and Observed Peak Ozone Concentrations (Denver, Hours 1100-1200 to 1400-1500 Local Standard Time, 28 July 1976) VI-52

VI-9 Possible Relationships Between the Model Performance Standards and a Guidelines Document VI-54

C-1 Locations and Values of Predicted Maximum One-Hour-Average Ozone Concentrations for Each Hour from 8 a.m. to 6 p.m. C-7

C-2 Concentration Histories Revealing Time Lag or Spatial Offset . . C-14

C-3 Estimate of Bias in Model Predictions as a Function of Ozone Concentration C-15

C-4 Time Variation of Differences Between Means of Observed and Predicted Ozone Concentrations C-17

C-5 Probabilities of Ozone Concentration Exceedance C-18

C-6 Model Predictions Correlated with Instrument Observations of Ozone (Data for 3 Days, 9 Stations, Daylight Hours) C-19

C-7 Model Predictions Compared with Estimates of Instrument Errors for Ozone (Data for 3 Days, 9 Stations, Daylight Hours) C-21

C-8 Map of Denver Air Quality Modeling Region Showing Air Quality Monitoring Stations C-23

C-9 Time History of Predicted and Observed Concentrations at Monitoring Sites C-24

C-10 Variations over All Stations of Observed and Predicted Average Ozone Concentrations C-25

C-11 Plots of Residuals and Forcing Variable C-26

C-12 Distribution of Area Fraction Exposed to Greater Than a Given Concentration Value C-30

C-13 Isopleths of Ozone Concentrations (pphm) on 29 July 1975 . . . . C-35

C-14 Size of Area in Which Predicted Ozone Concentrations Exceed Given Values for Years 1976, 1985, and 2000 . . . . C-40

C-15 Typical Residuals Isopleth Plot for Annual Average NO_2 . . . . . C-42

C-16 Estimated Exposure to Ozone as a Function of Ozone Concentration for 3 August 1976 Meteorology C-48

C-17 General Shape of the Exposure Cumulative Distribution and Density Functions C-49

C-18 Shape of $\psi(C)$, the Approximation to the Delta Function C-52

C-19 Cumulative Ozone Dosage as a Function of the Time of Day for 3 August 1976 Meteorology C-54

C-20 Cumulative Exposure (in 10^3 Person-Hours) to Ozone Concentrations Above Given Level in One-Square-Mile Grid Cells Between 500 and 1800 Hours for 3 August 1976 Meteorology and 1976 Emissions C-55

C-21 Cumulative Ozone Dosages (in 10^6 pphm-Person-Hours) in the One-Square-Mile Grid Cells from 500 to 1800 Hours (MST) for 3 August 1976 Meteorology and Emissions in 1976 . . . . C-58

C-22 Orientation with Respect to Measurement Station of Nearest Point at Which Prediction Equals Station Observation C-59

C-23 Space-Time Trace of Location of Nearest Point Predicting a Concentration Equal to the Station Measured Value C-60

D-1 Possible Health Effects Curves D-4

D-2 Representation of Spatial and Concentration Dependent Population Functions D-6

D-3 Population Distribution as a Function of Concentrations D-10

D-4 Idealized Concentration Isopleths D-11

D-5 Typical Radial Concentration Distributions About the Peak D-13

D-6 Predicted Population Distribution as a Function of Concentration D-16

D-7 Shifts in $\overline{w}(C)$ Caused by Nonuniform Population Distributions . . D-17

D-8 Expected Shape of Health Effects Function D-20

D-9 Minimum Allowable Ratio of Predicted to Measured Peak Concentration Value D-23

D-10 Prototypical Isopleth Diagram D-28

D-11 The Isopleth Diagram Replotted D-29

D-12 Total Regional Control Cost as a Function of the Level of Control Required D-32

D-13 Uncertainty Distribution for a Conservative Model D-35

D-14 Uncertainty Distribution for a Nonconservative Model D-35

TABLES

II-1 Air Quality Issues Commonly Addressed, by Generic Model Type II-5

II-2 Model/Application Combinations II-6

II-3 Some Air Quality Models II-7

II-4 Generic Performance Measure Information Requirements II-10

II-5 Types of Variations Among Generic Performance Measure Categories II-12

II-6 Performance Measures Commonly Associated with Specific Issues II-13

II-7 Performance Measures That Can Be Calculated by Each Model Type II-13

II-8 Performance Measure Objectives II-15

II-9 Importance of Performance Attributes by Issue II-16

II-10 Importance of Performance Attributes by Pollutant and Averaging Time II-18

II-11 Measures Recommended for Use in Setting Model Performance Standards II-19

II-12 Possible Rationales for Setting Model Performance Standards II-21

II-13 Performance Attributes Addressable Using Performance Standard Rationales II-22

II-14 Association of Rationales with Generic Issues II-22

II-15 Recommended Rationales for Setting Standards II-26

II-16 Summary of Recommended Performance Measures and Standards . . II-26

IV-1 Air Quality Issues Commonly Addressed, by Generic Model Type IV-18

IV-2 Possible Designations of Application Attributes IV-23

IV-3 Model/Application Combinations IV-24

IV-4 Some Air Quality Models IV-26

V-1 Generic Performance Measure Information Requirements V-8

V-2 Types of Variations Among Generic Performance Measure Categories V-11

V-3 Some Peak Performance Measures V-23

V-4 Some Station Performance Measures V-24

V-5 Some Area Performance Measures V-26

V-6 Some Exposure/Dosage Performance Measures V-28

V-7 Performance Measures Associated with Specific Issues V-34

V-8 Performance Measures That Can Be Calculated by Each Model Type V-37

VI-1 Performance Measure Objectives VI-10

VI-2 Importance of Performance Attributes by Issue. VI-10

VI-3 Importance of Performance Attributes by Pollutant and Averaging Time VI-13

VI-4 Candidate Station Performance Measures VI-16

VI-5 Useful Hybrid Performance Measures VI-20

VI-6 Measures Recommended for Use in Setting Model Performance Standards VI-21

VI-7 Possible Rationales for Setting Model Performance Standards . VI-24

VI-8 Performance Attributes Addressable Using Performance Standard Rationales VI-26

VI-9 Association of Rationales with Generic Issues VI-27

VI-10 Recommended Rationales for Setting Standards VI-29

VI-11 Summary of Recommended Performance Measures and Standards . . VI-31

VI-12 Sample Values for Model Performance Standards (Denver Example) VI-43

VI-13 Importance of Performance Attributes by Issue VI-56

VI-14 Importance of Performance Attributes by Pollutant and Averaging Time VI-56

VI-15 Model Performance Measures and Standards VI-57

B-1 Some Specific Air Quality Models B-3

C-1 Some Peak Performance Measures C-3

C-2 Several Peak Measure Combinations of Interest and Some Possible Interpretations C-4

C-3 Some Station Performance Measures C-8

C-4 Occurrence of Correspondence Levels of Predicted and Observed Ozone Concentrations C-20

C-5 Some Area Performance Measures C-29

C-6 Some Exposure/Dosage Performance Measures C-45

D-1 Selected Parameter Values in Denver Test Case D-15

EXHIBITS

III-1 Formal Organization of CFR Title 40--Protection of Environment III-4

IV-1 General Model Categories IV-3

I INTRODUCTION

In this report a candidate framework is suggested within which an objective evaluation of air quality simulation model (AQSM) performance may be carried out, along with an assessment of the relative applicability of models to specific problems. Quantitative procedures are identified that could facilitate assessment of the relative accuracy and usability of an AQSM.

The subject addressed in this report is a broad and complex one. Seldom can a rule for judging model performance be stated that does not have several plausible exceptions to it. Consequently, we view the establishment of model performance standards to be a pragmatic and evolutionary exercise. As we gain experience in evaluating model performance, we will need to modify both our choice of performance measures and the range of acceptable values we insist on. Nevertheless, the process must begin somewhere. The recommendations contained in this report represent such a beginning.

Model performance evaluation should not be viewed as a mechanistic process, to be performed in a "cookbook" fashion. Performance measures may be defined to be specific quantities whose value in some way characterizes the difference between predicted and observed concentrations. No set of performance measures, however well designed, can fully characterize model behavior. Judgment is required of the model user. Predictions can be compared with measurement data in a variety of ways. Some comparisons involve the calculation of specific quantities and are thus suited for having specific standards set. (An example might be the difference between the predicted and the observed concentration peak.) Other comparisons are more qualitative, better used in an advisory sense to facilitate "pattern recognition." (Concentration isopleth maps and time profiles of predicted and observed concentrations are examples of this type of qualitative comparison.) Although we recommend a set of performance measures and standards in this report, in no way does this recommendation suggest that computation of measures be limited to this set. For this reason, we catalogue many different types of performance measures, only a small subset of which have explicit, formal standards.

The measures and standards we suggest for use will almost certainly change as experience improves our "collective judgment" about what constitutes model acceptability and what does not. Perhaps the number of measures will increase to provide richer insight into model performance, or perhaps the number will shrink without any loss of "information content." Regardless of the list of measures and these standards that ultimately emerges for use, it is the conceptual structuring of the performance evaluation itself that seems to be most important at this point. We must identify clearly the desirable model attributes whose presence we are most interested in detecting, and we need to understand how we assess their relative importance, depending on the issue we are addressing and the pollutant species we are considering. This report offers a conceptual structure for "folding in" all these concerns and suggests candidate measures and standards.

A. OVERVIEW OF THE PROBLEM

Air quality simulation models (AQSMs) are widely used as predictive tools, estimating the impact on future air quality of alternative public decisions. Their predictions, however, are inherently nonverifiable. Only after the proposed action has been taken and the required implementation time elapsed will measurement data confirm or refute the model's predictive ability.

Herein lies the dilemma faced by users of air quality models: If a model's predictions at some future time cannot be verified, on what basis can we rely on that model to decide among policy alternatives? In resolving this dilemma, most users have adopted a pragmatic approach: If a model can demonstrate its ability to reproduce a set of "known" results for a similar type of application, then it is judged an acceptable predictive tool. It is on this basis that model "verification" has become an essential prelude to most modeling exercises.

Several investigators (Calder, 1974, and Johnson, 1972) have objected to this approach, arguing that it amounts to little more than "crude calibration." They suggest that true model validation can only be accomplished by evaluating each component sub-model--emissions, transport, or chemistry, for example. While this may be a scientifically sound approach, there are so many models available that it is difficult to complete such efforts for them all. Worse, the demand for a model, truly validated or not, often forces such concerns to be swept aside. We take a highly pragmatic position in this report, one that is also consistent with recommendations recently made to and by the U.S. Environmental Protection Agency [EPA] (Roth, 1977, and EPA, 1977). Because verification is so often performed at the "output end" (that is, only model results are examined, comparing them with "true" data), a systematic and objective procedure is needed in assessing model performance on that same basis.

A further difficulty exists. What constitutes a set of "known" results? This is not a problem easily solved. For "answers" to be known exactly, the "test" problem must be simple enough to be solved analytically. Few problems involving atmospheric dynamics are so simple. Most are complex and nonlinear. For these, the analytic test problem is an unacceptable one. Another, more practical alternative often is employed. For regional, multiple-source applications, the "known" results are taken to be the station measurements of concentrations actually recorded on a "test" date.

For source-specific applications, the source of interest may not yet exist, permission for its construction being the principal issue at hand. For these applications, it is often necessary to verify a model using the most appropriate of several prototypical "test cases." Though not existing currently, these could be assembled from measurements taken at existing sources, the variety of source size, type and location spanning the range of values found in applications of interest.

The term "known" is used imprecisely when referring to a set of measurement data. Station observations are subject to instrumentation error. The locations of fixed monitoring sites may not be sufficiently well distributed spatially to record data fully characterizing the concentration field and its peak value. Nevertheless, despite those shortcomings, "observed" data often are regarded as "true" data for the purpose of model verification.

In evaluating model performance, we must decide which performance attributes we most wish the model to possess. Having assembled two sets of data, one "known" and the other "predicted," we can assess model performance by comparing one with the other. Prediction and observation, however, can be compared in many ways. We must select the quantities (performance measures) that can most effectively test for the presence of those attributes.

Once we have decided on the performance measures best suited to our needs (and most feasible computationally), we can calculate these values. Having done so, however, we must ask a central question: How close must prediction be to observation in order for us to judge model performance as acceptable? If we are to answer "how good is good," performance standards for these measures must be set, with allowable tolerances (predicted values minus observed ones) derived from a reasonable rationale (health effects or pollution control cost considerations, for instance).

By setting these standards explicitly, certain benefits may be gained. Among these are the following:

- A degree of uniformity is introduced in assessing model reliability.
- A rational and objective basis is provided for comparing alternative models.
- The impact of limitations in both data gathering procedures and measurement network design can be made more explicit, facilitating any review of them that may be required.

> The performance expected of a model is stated clearly, in advance of the expenditure of substantial analysis funds, allowing model selection to be a more straightforward and less "risky" process.

> The needs for additional research can be identified clearly, with such efforts more directed in purpose.

B. STRUCTURE OF THE REPORT

The central purpose of this report is to suggest means for setting performance standards for air quality dispersion models. In doing so, our discussion proceeds in two phases, the first exploring key elements of the overall problem, as well as their interactions, and the second synthesizing all into a conceptual framework for model performance evaluation.

We recognize three key elements of the performance assessment problem, all of which are interrelated: the classes of issues addressed by AQSMs (air quality maintenance planning or prevention of significant deterioration, for example), the types of AQSMs available for use (grid-based, trajectory, or Gaussian models, for instance) with the applications for which they are suitable, and the classes of performance measures that are candidates for our use (two of which are station and exposure/dosage measures).

We consider each of these three elements in Chapters III, IV, and V, providing supporting material in Appendices A, B, and C. In Chapter III, we identify from current federal law and regulations seven distinctly different types of air quality issues, each of which may be addressed using an AQSM. In Chapter IV, we assess major model classes, examining their capabilities and limitations as well as their suitability for use in addressing each of the generic classes of issues. In Chapter V, we discuss model performance measures, identifying four major types, which we then assess for computational feasibility and suitability for use.

We provide supplementary detail for these three chapters in the first three appendicies. In Appendix A, we outline important portions of the Code of Federal Regulations. In Appendix B, we describe in summary form a number of specific air quality models. In Appendix C, we examine at length a variety of specific model performance measures, discussing their computation and providing illustrative examples of their calculation.

Having identified issues (Chapter III), issue/model combinations (Chapter IV), and issue/model/measure associations (Chapter V), we reach the synthesis phase in Chapter VI. Here we first identify five desirable attributes of model performance. Then we recommend a set of performance measures suitable for use in determining the presence or absence of each attribute. Each measure is chosen based on two criteria: First, it is an accurate indicator of the presence of a problem type and second, it is quantitative (that is, amenable to having specific standards set).

Having selected the performance measures for use, we then offer several possible rationales for determining the range of their acceptable values. We examine four rationales, discussing each in detail in Appendix D. Having done so, we recommend standards for use.

We also consider the way in which the relative importance of the five model performance attributes varies with the issue being addressed and the pollutant being considered. We recommend a means for ranking problem types that is dependent on these factors, using it as a way to decide from among procedural alternatives when a model fails to display a particular attribute.

To illustrate how to interpret the values of the recommended perforfance measures, we discuss a sample case. The sample case history is based on the use of the grid-based SAI Airshed Model in modeling the Denver Metropolitan region. Supplementary means for gaining insight into model behavior are also shown.

Finally, a conceptual framework is suggested for a draft model performance standard. The elements it should contain are discussed, as well as its relationship to a supplementary guidelines document.

With this final discussion, our presentation is complete, though the subject itself is by no means exhausted. Considerable additional effort is warranted, given the importance of this complex and difficult topic. We suggest in Chapter VII several areas in which we feel such work would prove fruitful.

II SUMMARY

In this chapter we summarize the results of this study. First, we state them in overall terms. Then, we summarize detailed results on a chapter-by-chapter basis.

A. MAIN RESULTS

Several main tasks are accomplished in this report. These represent the chief results of the study. We summarize them as follows:

> A conceptual framework is set for objective evaluation of dispersion model performance (Chapter VI).

> An outline for a draft model performance standards document is suggested (Chapter VI).

> Specific measures are recommended for use (Chapter VI).

> Specific rationales on which standards could be based are developed, several of which represent research that is original with this study (Chapter VI and Appendix D).

> Comprehensive background material is presented on key elements of the performance evaluation problem: the types of issues to be addressed (Chapter III and Appendix A), the classes of models to be used along with the applications for which they are suited (Chapter IV and Appendix B), and the categories of performance measures available for consideration (Chapter V and Appendix C).

> Guidance on the interpretation of performance measure values is provided by means of an illustrative sample case (Chapter VI).

B. DETAILED SUMMARY

Discussion in this report proceeds in two phases. In the first of these, we present a comprehensive examination of key elements of the performance evaluation problem. This background phase consists of the in-depth analysis in Chapters III, IV and V, supported by material in Appendices A, B and C.

We intend the background phase of this report to be regarded not as a supplement but rather as an essential prelude to the second, or synthesis, phase. The second phase, contained in Chapter VI and Appendix D, draws from the background material to identify a set of performance criteria that is both useful and computationally feasible.

In this section we present detailed summaries of the important results of the report. We do so on a chapter-by-chapter basis.

1. Summary of Chapter III (Issues)

This chapter provides an issues framework within which the application of air pollution models can be viewed. First, an overview is provided, highlighting important aspects of federal air pollution law (also see Appendix A). By means of this discussion, seven generic classes of issues are identified. These issues are examined and their implications for model applications explored.

The seven issue classes, divided into multiple-source and single-source categories, are described as follows:

> Multiple-Source Issues
  - SIP/C (State Implementation Plan/Compliance). The attainment of regional compliance with NAAQS, as considered in the SIP.
  - AQMP (Air Quality Maintenance Planning). Regional maintenance of compliance with the NAAQS, as considered in the SIP.

> Single-Source Issues

- PSD (Prevention of Significant Deterioration). Limitation of the amount by which the air quality may be degraded in areas in attainment of the NAAQS; this is considered in each SIP.
- NSR (New Source Review). Permit process by which applicants proposing new or modified stationary sources must demonstrate that both directly and indirectly caused emissions are within certain limits and that the pollution control to be employed is performed with the best available technology; this is considered in each SIP.
- OSR (Offset Rules). Interpretive decision by which all new or modified stationary sources in urban areas currently in noncompliance with the NAAQS are judged unacceptable unless the applicant can demonstrate a plan for reducing emissions in an existing source by an amount greater than the emissions from the proposed new sources; this decision has a strong impact on the stationary source permit process.
- EIS/R (Environmental Impact Statement/Report). A statement of impact required for major projects undertaken by the federal government or financed by federal funds (EIS), or a report of project impact required of public or private agencies by state or local statutes (EIR).
- LIT (Litigation). Court suits brought to resolve disagreement over any of the issues mentioned above or to secure variances waiving federal, state or local requirements.

2. Summary of Chapter IV (Models)

In Chapter III, we identified a set of generic air quality issues. In this chapter, we define a set of generic model types. Having done so, we match the two, identifying in generic terms those issues for which each model may be a suitable analysis tool. We also describe the technical formulations and underlying assumptions employed in each generic model type, indicating some key limitations. Through this presentation, we specify the relationship between generic issues, models, and the applications for which they are suitable.

The generic classes of dispersion models that we consider are:

- > Rollback
- > Isopleth
- > Physico-chemical
  - Grid
    - Region Oriented
    - Specific Source Oriented
  - Trajectory
    - Region Oriented
    - Specific Source Oriented
  - Gaussian
    - Long-Term Averaging
    - Short-Term Averaging
  - Box

In Table II-1 we associate generic model types with air quality issues for which their use is most appropriate. In Table II-2 we present model/application combinations of interest, characterizing applications by five attributes: number of sources, area type, pollutant, terrain complexity, and required resolution. The table lists the values of the attributes that can be accommodated by each model type.

In Table II-3 we relate some specific air quality models to the generic model categories in which they may be classified. Each of these models is described in detailed summary form in Appendix B.

3. Summary of Chapter V (Performance Measures)

In this chapter we discuss the types of performance measures available for use, examining their relationship with both the issues

TABLE II-1. AIR QUALITY ISSUES COMMONLY ADDRESSED BY GENERIC MODEL TYPE

Generic Model Type	Issue Category						
	SIP/C	AQMP	PSD	NSR	OSR	EIS/R	LIT
Refined Usage							
1. Grid[1]							
a. Region Oriented	X	X	X	X^2	X	X	X
b. Specific Source Oriented			X	X	X^3	X	X
2. Trajectory[1]							
a. Region Oriented	X	X			X	X	X
b. Specific Source Oriented			X	X	X^3	X	X
3. Gaussian[3]							
a. Short-Term Averaging[1]							
i) Multiple Source	X	X	X		X	X	X
ii) Single Source	X		X	X	X	X	X
b. Long Term-Averaging[4]	X	X	X	X	X	X	X
Refined/Screening Usage							
4. Isopleth[1,5]	X	X					
Screening Usage							
5. Rollback	X	X					
6. Box	X	X					

Notes:

1. Only short-term time scales can be considered (less than several days).
2. Regional impact of new sources can be assessed but not near-source, or microscale, effects.
3. Only non-reactive pollutants can be considered.
4. Only pollutants having long-term standards can be considered (SO_2, TSP, and NO_2).
5. Only photochemically active pollutants can be considered.

TABLE II-2. MODEL/APPLICATION COMBINATIONS

Generic Model Type	Number of Sources	Area Type	Pollutant	Terrain Complexity	Required Resolution
REFINED USAGE					
Grid					
a. Region Oriented	Multiple-Source	Urban Rural	O_3, HC, CO, NO_2 (1-hour), SO_2 (3- and 24-hour), TSP	Simple Complex (Limited)	Temporal Spatial
b. Specific Source Oriented	Single-Source	Rural	O_3, HC, CO, NO_2 (1-hour), SO_2 (3- and 24-hour), TSP	Simple Complex (Limited)	Temporal
Trajectory					
a. Region Oriented	Multiple-Source	Urban	O_3, HC, CO, NO_2 (1-hour), SO_2, (3- and 24-hour), TSP	Simple	Temporal Spatial (Limited)
b. Specific Source Oriented	Single-Source	Urban Rural	O_3, HC, CO, NO_2 (1-hour), SO_2, (3- and 24-hour), TSP	Simple Complex (Limited)	Temporal Spatial (Limited)
Gaussian					
a. Long-Term Averaging	Multiple-Source Single-Source	Urban Rural	SO_2 (Annual), TSP, NO_2 (Annual)*	Simple Complex (Limited)	Spatial
b. Short-Term Averaging	Multiple-Source Single-Source	Urban Rural	SO_2 (3- and 24-hour), CO, TSP, NO_2, (1-hour)*	Simple Complex (Limited)	Temporal Spatial
REFINED/SCREENING USAGE					
Isopleth	Multiple-Source	Urban	O_3, HC, NO_2 (1-hour)	Simple Complex (Limited)	Temporal (Limited)
SCREENING USAGE					
Rollback	Multiple-Source Single-Source	Urban Rural	O_3, HC, NO_2 SO_2, CO, TSP	Simple Complex (Limited)	--
Box	Multiple-Source	Urban	O_3, HC, CO, NO_2 (1-hour), SO_2 (3- and 24-hour), TSP	Simple Complex (Limited)	Temporal

---

* Only if NO_2 is taken to be total NO_x.

TABLE II-3. SOME AIR QUALITY MODELS

Generic Model Type	Specific Model Name
Refined Usage	
Grid	
a. Region Oriented	SAI LIRAQ PICK
b. Specific Source Oriented	EGAMA DEPICT
Trajectory	
a. Region Oriented	DIFKIN REM ARTSIM
b. Specific Source Oriented	RPM LAPS
Gaussian	
a. Long-term Averaging	AQDM CDM CDMQC TCM ERTAQ* CRSTER* VALLEY* TAPAS*
b. Short-term Averaging	APRAC-1A CRSTER* HANNA-GIFFORD HIWAY PTMTP PTDIS PTMAX RAM VALLEY* TEM TAPAS* AQSTM CALINE-2 ERTAQ*
Refiner/Screening Usage	
Isopleth	EKMA WHITTEN
Screening Usage	
Rollback	LINEAR ROLLBACK MODIFIED ROLLBACK APPENDIX J
Box	ATDL

---

* These models can be used for both long-term and short-term averaging.

and the models we identified in Chapters III and IV. Our discussion proceeds as follows: We first identify generic types of performance measures; we then catalogue some specific performance measures (describing them in detail in Appendix C); and finally we match generic performance measures to the issue/model/application combinations presented in earlier chapters.

We consider four generic performance measure categories: peak, station, area, and exposure/dosage. The first category contains those measures deriving from the differences between the predicted and observed concentration peak, its level, location and timing. The second category includes measures based on concentration differences between prediction and observation at specific measurement stations. Within the third category are contained those measures based on concentration field differences throughout a specified area. The fourth category includes measures derived from differences in population exposure and dosage within a specified area.

Each of these generic performance measure categories requires successively greater knowledge of the spatial and temporal distribution of concentrations. We show in Figure II-1 a schematic representation of several distinct levels of knowledge about regional concentrations. A similar schematic illustration appropriate for source-specific situations is shown in Figure II-2. Listed in Table II-4 are the information requirements for the four categories. We also consider the relative likelihoods that reliable information will be available supporting calculation of measures from each of the four categories.

Three types of variations are recognized among performance measures: scalar, statistical, and pattern recognition. Those measures of the first type are based on a comparison of the predicted and observed values of a specific quantity: the peak concentration level, for instance. Those of the second type compare the statistical behavior (the mean, variance, and correlation, for example) of the differences between the predicted and observed values for the quantities of interest.

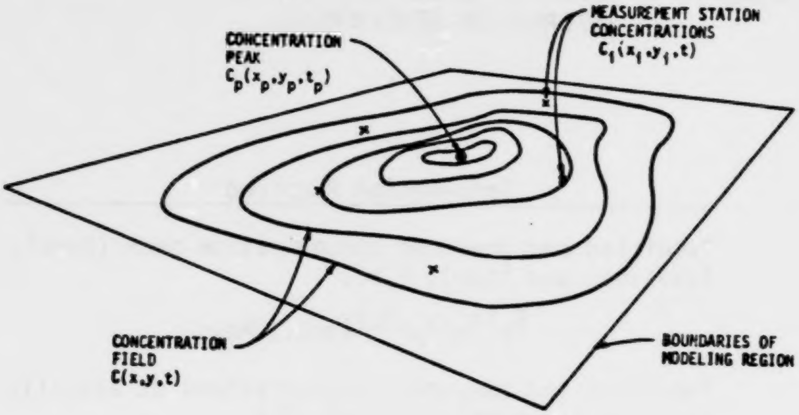

CONCENTRATION PEAK
C_p(x_p,y_p,t_p)
MEASUREMENT STATION CONCENTRATIONS
C_i(x_i,y_i,t)
CONCENTRATION FIELD
C(x,y,t)
BOUNDARIES OF MODELING REGION

FIGURE II-1. VARIOUS LEVELS OF KNOWLEDGE ABOUT REGIONAL CONCENTRATIONS

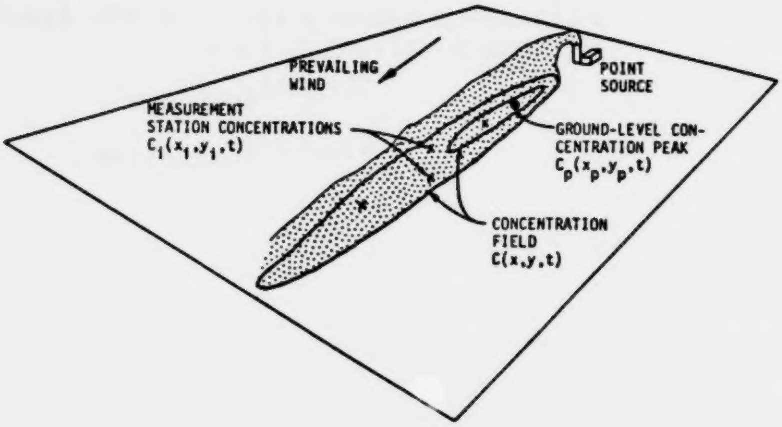

PREVAILING WIND
POINT SOURCE
MEASUREMENT STATION CONCENTRATIONS
C_i(x_i,y_i,t)
GROUND-LEVEL CONCENTRATION PEAK
C_p(x_p,y_p,t)
CONCENTRATION FIELD
C(x,y,t)

FIGURE II-2. VARIOUS LEVELS OF KNOWLEDGE ABOUT SPECIFIC-SOURCE CONCENTRATIONS

TABLE II-4. GENERIC PERFORMANCE MEASURE INFORMATION REQUIREMENTS

Generic Performance Measure Type	Information Required
Peak	Predicted and measured concentration peak (level, location, and time), i.e., $C_p(x_p,y_p,t_p)_{Pred., Meas.}$
Station	Predicted and measured concentrations at specific stations (temporal history), i.e., $C_i(x_i,y_i,t)_{Pred.,Meas.}$, $1 \leq i \leq N$ stations
Area	Predicted and measured concentration field within a specified area (spatial and temporal history), i.e., $C(x,y,t)_{Pred.,Meas.}$
Exposure/dosage	Both the predicted and measured concentration field and the predicted and actual population distribution within a specified area (spatial and temporal history), i.e., $C(x,y,t)_{Pred.,Meas.}$ $C(x,y,t)_{Pred.,Actual}$

Measures of the final type are useful in triggering "pattern recognition," that is, providing qualitative insight into model behavior, transforming concentration "residuals" (the differences between predicted and observed values) into forms that highlight certain aspects of model performance.

To illustrate the types of variations found in each generic performance measure category, we present Table II-5. Some typical examples are included for each category/variation combination. In Section D of this chapter, a number of specific performance measures are listed. Examined in detail in Appendix C, they are classified according to the scheme presented here.

For reasons we examine in this chapter, performance measures may be associated with the issue classes. We match issue with measure in Table II-6, indicating where their calculation might be of use. Note that NSR and PSD are both part of the preconstruction review process for a new source.

Also, we may match measures to model type, as is shown in Table II-7. This we do based on differences among model types in their ability to calculate each of the measure types. Isopleth, rollback and box models, for instance, provide insufficient spatial resolution for calculation of station, area or exposure/dosage measures. Likewise, long-term averaging Gaussian models lack sufficient temporal resolution to permit calculation of exposure/dosage measures.

Several important conclusions are reached in this chapter about the suitability for use of each of the four measure types:

> Performance measures relying on a comparison of the predicted and "true" peak concentrations may not be reliable in all circumstances, since measurement networks can provide only the concentration at the station recording the highest value, not necessarily the value at the "true" peak.

TABLE II-5. TYPES OF VARIATIONS AMONG GENERIC PERFORMANCE MEASURE CATEGORIES

Generic Performance Measure Category	Types of Variations	Typical Example
Peak	Scalar	Concentration residual* at the peak.
	Pattern Recognition	Map showing locations and values of maximum one-hour-average concentrations for each hour.
Station	Scalar	Concentration residual at the station measuring the highest value.
	Statistical	Expected value, variance and correlation coefficient of the residuals for the modeling day at a particular measurement station.
	Pattern Recognition	At the time of the peak (event-related), the ratio of the residual at the station having the highest value to the average of the residuals at the other station sites (this can indicate whether the model performs better near the peak than it does throughout the rest of the modeled region).
Area	Scalar	Difference in the fraction of the modeled area in which the NAAQS are exceeded.
	Statistical	At the time of the peak, differences in the area/concentration frequency distribution.
	Pattern Recognition	For each modeled hour, isopleth plots of the ground-level residual field.
Exposure/dosage	Scalar	Differences in the number of person-hours of exposure to concentrations greater than the NAAQS.
	Statistical	Differences in the exposure concentration frequency distribution.
	Pattern Recognition	For the entire modeled day, an isopleth plot of the ground level dosage residuals.

---

* Residual: The difference between "predicted" and "observed."

TABLE II-6. PERFORMANCE MEASURES COMMONLY ASSOCIATED WITH SPECIFIC ISSUES

	Performance Measure Type			
Issue	**Peak**	**Station**	**Area**	**Exposure/Dosage**
Multiple-source				
SIP/C	X	X	X	X
AQMP	X	X	X	
Specific-source				
PSD	X	X	X	
NSR	X	X	X	
DSR		X	X	
EIS/R	X	X	X	X
LIT	X	X	X	

TABLE II-7. PERFORMANCE MEASURES THAT CAN BE CALCULATED BY EACH MODEL TYPE

	Performance Measure Type			
Model	**Peak**	**Station**	**Area**	**Exposure/ Dosage**
Refined usage				
Grid				
Region oriented	X	X	X	X
Specific source oriented	X	X	X	X
Trajectory				
Region oriented	X	X		
Specific source oriented	X	X	X	X
Gaussian				
Long-term averaging	X	X	X	
Short-term averaging	X	X	X	X
Refined/screening usage				
Isopleth	X			
Screening usage				
Rollback	X			
Box	X			

- > Performance measures relying on a comparison of the predicted and "true" concentration fields may not be computationally feasible since neither predicted nor "true" concentration fields are always resolvable, spatially or temporally.
- > Performance measures based upon a comparison of predicted and "true" exposure/dosage, though they are appealing because of their ability to serve as surrogates for the health effects experienced by the populace, may not be computationally feasible because of the difficulty in measuring the "true" population distribution and the "true" concentration field. (We do suggest in Chapter VI and Appendix D, however, one means by which health effects considerations can be accounted for implicitly.)
- > Performance measures based upon a comparison of the predicted and observed concentrations at station sites in the measurement network may be of the greatest practical value.

4. Summary of Chapter VI (Performance Standards)

The central purpose of this report is to suggest means for setting performance standards for air quality dispersion models. In this chapter we reach this goal. Our discussion proceeds as follows: First we identify five key attributes of desirable model performance, evaluating how their relative importance varies depending on the issue addressed and the pollutant/averaging time considered; then we propose specific performance measures appropriate for use in testing for the presence of these attributes; and finally we suggest rationales on which to base the setting of formal standards. Having recommended for use a list of performance measures and standards, we deal with two additional issues: interpretation of the values of the measures, which we illustrate by means of a sample case study, and promulgation of formal performance criteria, which we explore by proposing an outline of a draft standard.

The five attributes of desirable model performance are defined as follows: accuracy of the peak prediction, absence of systematic bias, lack of gross error, temporal correlation, and spatial alignment. Though they are interrelated, each of the five performance attributes is distinct. Consequently, we must employ different kinds of performance measures to determine the presence or absence of each. We list in Table II-8 the objectives of each type of performance measure.

TABLE II-8. PERFORMANCE MEASURE OBJECTIVES

Performance Attributes	Objective of Performance Measures
Accuracy of the peak prediction	Assess the model's ability to predict the concentration peak (its level, timing and location)
Absence of systematic bias	Reveal any systematic bias in model predictions
Lack of gross error	Characterize the error in model predictions both at specific monitoring stations and overall
Temporal correlation	Determine differences between predicted and observed temporal behavior
Spatial alignment	Uncover spatial misalignment between the predicted and observed concentration fields

We classify the difference between bias and error by means of the following example. Suppose when we compare a set of model predictions with station observations, we find several large positive residuals (predicted minus observed concentrations) balanced by several equally large negative residuals. If we were testing for bias, we would allow the oppositely signed residuals to cancel. A conclusion that the model displayed no systematic bias therefore might be a justifiable one. On the other hand, were we testing for gross error, the signs of the residuals would not be considered with oppositely signed residuals no longer allowed to cancel. Because the absolute value of the residuals is large in our example, we might well conclude that the model predictions are subject to significant gross error.

Which of these performance attributes, however, is most important? This question has no unique answer, the relative importance of each attribute depending on the type of issue the model is being used to address and the type of pollutant under consideration. In order to relate attribute importance to application issue in a more convenient manner, we present in Table II-9 a matrix of generic issues (as defined earlier in this report) and problem type. For each combination we indicate an "importance category." We define the three categories based on how strongly we insist that model performance be judged acceptable for the given problem type. For Category 1, we require that the performance attribute must be present (the problem type is of prime importance). For Category 2, the attribute should be present but, if it is not, some leeway ought to be allowed, perhaps at the discretion of a reviewer (although the attribute is of considerable importance, some degree of "mismatch" may be tolerable). For Category 3, we are not insistent that the performance attribute be present, though we state that as being a desirable objective (the attribute is not of central importance). The reasoning behind the entries in this table is complex. For this reason, we urge the reader to consult the detailed discussion in Chapter IV Section C.

TABLE II-9. IMPORTANCE OF PERFORMANCE ATTRIBUTES BY ISSUE

	Importance of Performance Attribute*						
Performance Attribute	SIP/C	AQMP	PSD	NSR	OSR	EIS/R	LIT
Accuracy of the peak prediction	1	1	1	1	2	1	1
Absence of systematic bias	1	1	1	1	1	1	1
Lack of gross error	1	1	1	1	1	1	1
Temporal correlation	2	2	3	3	3	3	3
Spatial alignment	2	2	1	3	3	3	3

---

* Category 1 - Performance standard must always be satisfied.
Category 2 - Performance standard should be satisfied, but some leeway may be allowed at the discretion of a reviewer.
Category 3 - Meeting the performance standard is desirable but failure is not sufficient to reject the model; measures dealing with this problem should be regarded as "informational."

The relative importance of each performance attribute also is dependent on the type of pollutant being considered and the averaging time required by the NAAQS. If a species is subject to a short-term standard, for instance, model peak accuracy and temporal correlation might be of considerable concern, depending on the issue being addressed. However, if the species is subject to a long-term standard, neither of these are of appropriate form. We indicate in Table II-10 a matrix of the problem types and pollutant species. We rank each combination by the same importance categories we used earlier in Table II-3.

Conceivably, a conflict might exist between the ranking indicated by the issue and the pollutant matrices in Tables II-9 and II-10. We would resolve the conflict in favor of the less stringent of the two rankings.

Having identified the problem types of interest, we then suggest specific performance measures for use. Our recommended choice of performance measures is based upon the following criteria:

> The measure is an accurate indicator of the presence of a given problem type.
> The measure is of the "absolute" kind, that is, specific standards can be set.
> Only station measures should be considered for use in setting standards.* (This is more an "unavoidable" choice than a "preferred" one.)

Based on these criteria, we recommend the set of measures described in Table II-11. The use of ratios (C_{p_p}/C_{p_m} and $\bar{\mu}$, for example) can introduce difficulties: They can become unstable at low concentrations, and the statistics of a ratio of two random variables can become troublesome. Nevertheless, when used properly their advantages can be offsetting. For example, the use of C_{p_p}/C_{p_m} instead of C_p-C_m) permits a health effects rationale to be used in recommending a performance standard (see a later discussion).

*Note the caveat on pages VI-18 and VI-19, with respect to point source applications.

We draw a distinction between those measures that are of general use in examining model performance and the much smaller subset of measures that are most amenable to the establishment of explicit standards. Many measures can provide rich insight into model behavior, but the information is conveyed in a qualitative way not suitable for quantitative characterization (a requisite for use in setting performance standards).

TABLE II-10. IMPORTANCE OF PERFORMANCE ATTRIBUTES BY POLLUTANT AND AVERAGING TIME

	Importance of Performance Attribute*										
	Pollutants with Short-term Standards								Pollutants with Long-term Standards		
Performance Attribute	O_3** (1 hour)§	CO** (1 hour)	NMHC* (3 hour)	SO_2 (3 hour)	NO_2 (?)†	CO (8 hour)	TSP** (24 hour)	SO_2** (24 hour)	NO_2** (1 year)	TSP (1 year)	SO_2 (1 year)
Accuracy of the peak prediction	1	1	1	1	1	1	1	1	3	3	3
Absence of systematic bias	1	1	1	1	1	1	1	1	1	1	1
Lack of gross error	1	1	1	1	1	1	1	1	1	1	1
Temporal correlation	1	2	2	2	1	2	3	3	N/A††	N/A	N/A
Spatial alignment	1	2	2	2	1	2	2	2	2	2	2

---

* Category 1 - Performance standard must be satisfied.
Category 2 - Performance standard should be satisfied, but some leeway may be allowed at the discretion of a reviewer.
Category 3 - Meeting the performance standard is desirable but failure is not sufficient to reject the model.

† No short-term NO_2 standard currently exists.

§ Averaging times required by the NAAQS are in parentheses.

** Primary standards.

†† The performance attribute is not applicable.

TABLE II-11. MEASURES RECOMMENDED FOR USE IN SETTING MODEL PERFORMANCE STANDARDS†

Performance Attribute	Performance Measure
Accuracy of the peak prediction	Ratio of the predicted station peak to the measured station (could be at different stations and times) C_{p_p}/C_{p_m} Difference in timing of occurrence of station peak* Δt_p
Absence of systematic bias	Average value and standard deviation of the mean deviation about the perfect correlation line normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either the predicted or the observed values exceed some appropriate minimum value (possibly the NAAQS) $\left(\bar{\mu}, \sigma_{\bar{\mu}}\right)_{\mathrm{OVERALL}}$
Lack of gross error	Average value and standard deviation of the absolute deviation about the perfect correlation line normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either the predicted or the observed values exceed some appropriate minimum value (possibly the NAAQS) $\left(\lvert\bar{\mu}\rvert, \sigma_{\lvert\bar{\mu}\rvert}\right)_{\mathrm{OVERALL}}$
Temporal correlation*	Temporal correlation coefficients at each monitoring station for the entire modeling period and an overall coefficient averaged for all stations $r_{t_i}, r_{t_{\mathrm{OVERALL}}}$ for $1 \leq i \leq M$ monitoring stations
Spatial alignment	Spatial correlation coefficients calculated for each modeling hour considering all monitoring stations, as well as an overall coefficient average for the entire day $r_{x_j}, r_{x_{\mathrm{OVERALL}}}$ for $1 \leq j \leq N$ modeling hours

---

* These measures are appropriate when the chosen model is used to consider questions involving photochemically reactive pollutants subject to short-term standards.

† There is deliberate redundancy in the performance measures. For example, in testing for systematic bias, $\bar{\mu}$ and $\sigma_{\bar{\mu}}$ are calculated. The latter quantity is a measure of "scatter" about the perfect correlation line. This is also an indicator of gross error and could be used in conjunction with $\lvert\bar{\mu}\rvert$ and $\sigma_{\bar{\mu}}$.

These "measures," often involving graphical display, really are tools for use in "pattern recognition." They display model behavior in suggestive ways, highlighting "patterns" whose presence reveals much about model performance. Several examples of such "measures" are isopleth contour maps of predicted concentrations estimates of "observed" ones, isopleth contour maps of the differences between the two, and time histories of predicted and observed concentrations at specific monitoring stations.

Although we focus on station measures for use in setting model performance standards, we do not suggest that the calculation of performance measures be limited to such measures. Many other measures should be used where appropriate. The data should be viewed in as many, varied ways as possible in order to enrich insight into model behavior. We suggest a number of useful measures both in Chapter V and Appendix C.

Having identified specific measures for use, we consider four rationales for setting appropriate standards. The rationales, along with a statement of their guiding principles, are shown in Table II-12. We discuss each in detail in Appendix D.

The four rationales differ in their ability to consider each of the five problem types. Shown in Table II-13 are the types of problems addressable by measures whose standards are set by each of the rationales. Only the Pragmatic/Historic rationale is of use in addressing all problem types; the other three are of use principally in defining the level of performance required in predicting values at or near the concentration peak. In Table II-14 we associate each rationale with those issues for which its use is appropriate.

We select in the following ways from among the alternative rationales. Hoping to avoid introducing a procedural bias, we first eliminate the Guaranteed Compliance rationale from further consideration. Then, because the Health Effects rationale is better suited for use in setting

TABLE II-12. POSSIBLE RATIONALES FOR SETTING MODEL PERFORMANCE STANDARDS

Rationale	Guiding Principle
Health Effects	The metric of concern is the area-integrated cumulative health effects due to pollutant exposure; the ratio of the metric's value based on prediction to its value based on observation must be kept to within a prescribed tolerance of unity.
Control Level Uncertainty	Uncertainty in estimates of the percentages of emissions control required must be kept within certain allowable bounds.
Guaranteed Compliance	Compliance with the NAAQS must be "guaranteed"; all uncertainty must be on the conservative side even if this approach means introducing a systematic bias.
Pragmatic/Historic	In each new application, a model should perform at least as well as the "best" previous performance of a model in its generic class in a similar application; until such a historical data base is complete, other more heuristic approaches may be applied.

TABLE II-13. PERFORMANCE ATTRIBUTES ADDRESSABLE USING PERFORMANCE STANDARD RATIONALES

Performance Attribute	Health* Effects	Control Level* Uncertainty	Guaranteed Compliance	Pragmatic/ Historic
Accuracy of the peak prediction	X	X	X	X
Absence of systematic bias				X
Lack of gross error	X		X	X
Temporal correlation				X
Spatial alignment				X

---

* These are most suited for photochemically reactive pollutants subject to short-term standards.

TABLE II-14. ASSOCIATION OF RATIONALES WITH GENERIC ISSUES

	Issue Category						
	Multiple-Source		Specific-Source				
Rationale	SIP/C	AQMP	PSD	NSR	OSR	EIS/R	LIT
Health Effects	X	X	X	X		X	X
Control Level Uncertainty	X	X	X	X		X	X
Guaranteed Compliance	X	X	X	X		X	X
Pragmatic/ Historic	X	X	X	X	X	X	X

standards for peak measures, we choose to use it only in that way. As is clear from Table II-13, we presently have no alternative but to apply the Pragmatic/Historic rationale for those measures designed to test for systematic bias and gross error as well as to evaluate temporal correlation and spatial alignment.

Where we invoke the Pragmatic/Historic rationale as justification for selecting specific standards, we also state the specific guiding principles we follow. We summarize those here:

> When the pollutant being considered is subject to a short-term standard, the timing of the concentration peak may be an important quantity for a model to predict. This is particularly true when the pollutant is also photochemically reactive. We state as a guiding principle: "For photochemically reactive pollutants, the model must reproduce reasonably well the phasing of the peak." For ozone an acceptable tolerance for peak timing might be ±1 hour.

> The model should not exhibit any systematic bias at concentrations at or above some appropriate minimum value (possibly the NAAQS) greater than the maximum resulting from EPA-allowable calibration error in the air quality monitors. We would consider in our calculations any prediction-observation pair in which either of the values exceed the pollutant standard.

> Error (as measured by its mean and standard deviation) should not be significantly different from the distribution of differences resulting from the comparison of an EPA-acceptable monitor with an EPA reference monitor. The EPA has set maximum allowable limits on the amount by which a monitoring technique may differ from a reference method (40 CFR § 53.20). An "EPA-acceptable monitor" is defined here to be one that differs from a reference monitor by up to the maximum allowable amount.

> Predictions and observations should appear to be highly correlated at a 95 percent confidence level, both when compared

> temporally and spatially. We can estimate the minimum allowable value for the respective correlation coefficient by using a t-statistic at the appropriate percentage level and having the degrees of freedom appropriate for the number of prediction-observation pairs.

The guiding principles noted above are plausible ones, though in some cases they are arbitrary. As a "verification data base" of experience is assembled, historically achieved performance levels may be better indicators of the expected level of model performance. Standards derived on this more pragmatic basis may supplant those deriving from the "guiding principles" followed in this report.

Our recommended choice for use, when possible, in establishing peak-accuracy standards is a composite one, combining the Health Effects and Control Level Uncertainty rationales. Were a model to overpredict the peak, a control strategy based on its prediction might be expected to abate the health impact actually occurring, though with more control than actually needed. If the model underpredicted, however, the control strategy might be "underdesigned," with the risk existing that some of the health impact might remain unabated even after control implementation. The penalty, in a health sense, is incurred only when the model underpredicts. The Health Effects rationale then is one-sided, helping us set performance standards only on the "low side."

On the other hand, the Control Level Uncertainty rationale is bounded "above" and "below", that is, its use provides a tolerance interval about the value of the measured peak concentration. For a model to be judged acceptable under this criterion, its prediction of the peak concentration would have to fall within this interval. Model underprediction could lead to control levels lower than required, but residual health risks. Overprediction, on the other hand, could lead to abatement strategies posing little or no health risk but incurring control costs greater than required.

For the above reasons, we suggest that the Control Level Uncertainty rationale be used to establish an upper bound (overprediction) on the acceptable difference between the predicted and observed peak. We would choose the lower bound (underprediction) to be the interval that is the minimum of that suggested by the Health Effects and Control Level Uncertainty rationales.

We list our recommendations in Table II-15, noting the possibility that the recommended rationales may not be appropriate in all applications for all pollutants. Whether health effects would be an appropriate consideration when considering TSP, for instance, is unclear. The Health Effects rationale, as defined in Appendix D, is best suited for use in urban applications involving short-term, reactive pollutants. In those circumstances when the HE or CLU rationales are not suitable, we suggest the Pragmatic/Historic rationale.

We summarize in Table II-16 our list of recommended performance measures and standards. In it, we associate performance attribute and standard. To further describe the standard, we state the type of rationale used and the guiding principle followed, as well as providing sample values that are appropriate for the sample case we consider in this chapter.

We also discuss two supplementary subjects. First, we illustrate how performance measure values may be interpreted by describing a sample case based on use of the SAI Airshed Model in simulating the Denver Metropolitan region Then, we consider means by which model performance criteria may be promulgated, suggesting an outline for a draft standard.

Thus we conclude this chapter and the report. We note in closing that the performance subject itself is by no means exhausted. Many areas remain to be explored in greater detail, all warranting considerable additional effort.

TABLE II-15. RECOMMENDED RATIONALES FOR SETTING STANDARDS

Performance Attribute	Recommended Rationale
Accuracy of peak prediction	Health Effects* (lower side/underprediction) Control Level Uncertainty* (upper side/overprediction)
Absence of sytematic bias	Pragmatic/Historic
Lack of gross error	Pragmatic/Historic
Temporal correlation	Pragmatic/Historic
Spatial alignment	Pragmatic/Historic

---

* These may not be appropriate for all regulated pollutants in all applications. When they are not, the Pragmatic/Historic rationale should be employed. They are most applicable for photochemically reactive pollutants subject to a short-term standard (O_3 and NO_2, if a 1-hour standard is set).

TABLE II-16. SUMMARY OF RECOMMENDED PERFORMANCE MEASURES AND STANDARDS

Performance Attribute	Performance of Measure	Performance Standard		
		Type of Rationale	Guiding Principle	Sample Value (Denver Example)
Accuracy of the peak prediction	Ratio of the predicted station peak to the measured station peak (could be at different stations and times) C_{p_p}/C_{p_m}	Health Effects[†] (lower side) combined with Control Level Uncertainty (upper side)	Limitation on uncertainty in aggregate health impact and pollution abatement costs[†]	$80 \le \frac{C_{p_p}}{C_{p_m}} \le 150$ percent
	Difference in timing of occurrence of station peak[*] Δt_p	Pragmatic/Historic	Model must reproduce reasonably well the phasing of the peak, say, ±1 hour	± 1 hour
Absence of systematic bias	Average value and standard deviation of mean deviation about the perfect correlation line normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either predicted or observed values exceed some appropriate minimum value (possibly the NAAQS). $\left(\overline{u}, \sigma_{\overline{u}}\right)_{OVERALL}$	Pragmatic/Historic	No or very little systematic bias at concentrations (predictions or observations) at or above some appropriate minimum value (possibly the NAAQS); the bias should not be worse than the maximum bias resulting from EPA-allowable monitor calibration error (-8 percent is a representative value for ozone); the standard deviation should be less than or equal to that of the difference distribution of an EPA-acceptable monitor[**] compared with a reference monitor. (3 pphm is representative for ozone at the 95 percent confidence level)	No apparent bias at ozone concentrations above 0.06 ppm (see Table VI-12 and Figures VI-5 and VI-6 for further details)
Lack of gross error	Average value and standard deviation of absolute mean deviation about the perfect correlation line normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either predicted or observed values exceed some appropriate minimum value (possibly the NAAQS) $\left(\lvert\overline{u}\rvert, \sigma_{\lvert\overline{u}\rvert}\right)_{OVERALL}$	Pragmatic/Historic	For concentrations at or above some appropriate minimum value (possibly the NAAQS), the error (as measured by the overall values of $\lvert\overline{u}\rvert$ and $\sigma_{\lvert\overline{u}\rvert}$) should be indistinguishable from the difference resulting from comparison of an EPA-acceptable monitor with a reference monitor	NO excessive gross error (see Table VI-12 and Figures VI-5 and VI-6 for further details)
Temporal correlation[*]	Temporal correlation coefficients at each monitoring station for the entire modeling period and an overall coefficient for all stations $r_{t_i}, r_{t_{OVERALL}}$, for $1 \le i \le M$ monitoring stations	Pragmatic/Historic	At a 95 percent confidence level, the temporal profile of predicted and observed concentrations should appear to be in phase (in the absence of better information, a confidence interval may be converted into a minimum allowable correlation coefficient by using an appropriate t-statistic)	For each monitoring station, $0.69 \le r_{t_i} \le 0.97$ Overall, $r_{t_{OVERALL}} = 0.88$ In this example a value of $r \ge 0.53$ is significant at the 95 percent confidence level
Spatial alignment	Spatial correlation coefficients calculated for each modeling hour considering all monitoring stations, as well as an overall coefficient for the entire day $r_{x_j}, r_{x_{OVERALL}}$, for $1 \le j \le N$ modeling hours	Pragmatic/Historic	At a 95 percent confidence level, the spatial distribution of predicted and observed concentrations should appear to be correlated	For each hour, $-0.43 \le r_{x_j} \le 0.66$ Overall, $r_{x_{OVERALL}} = 0.17$ In this example a value of $r \ge 0.71$ is significant at the 95 percent confidence level

---

[*] These measures are appropriate when the chosen model is used to consider questions involving photochemically reactive pollutants subject to short-term standards.

[†] These may not be appropriate for all regulated pollutants in all applications. When they are not the Pragmatic/historic rationale should be employed.

[**] The EPA has set maximum allowable limits on the amount by which a monitoring technique may differ from a reference method. An "EPA-acceptable monitor" is defined here to be one that differs from a reference monitor by up to the maximum allowable amount.

III ISSUES REQUIRING MODEL APPLICATION

Air pollution models have been developed over a period of years, not always in response to specific needs. While convenience and availability (rather than strict suitability) often motivated their use in particular applications, certain classes of models have come to be associated with certain classes of applications. For this reason, it is helpful to view the setting of model performance measures and standards within that issue-specific context. This chapter is intended to provide an issues framework within which the application of air pollution models can be viewed. First, an overview is provided, highlighting important aspects of air pollution law. By means of this discussion, generic issues are identified. Then, these issues are examined and their implications for model applications explored.

A. A PERSPECTIVE ON THE ISSUES

Basic air pollution law in this country has been enacted at the federal level, although many important legal variants exist among states and localities. The passage of legislation, however, is often just a first step. Usually, only broad authority is granted in the original law. It remains to the federal agency thus chartered by the Congress to set the specific regulations implementing the law. These are then promulgated, becoming an additional part of the Code of Federal Regulations (CFR). Notice is provided of such an action by publication in the Federal Register (FR). When disagreements exist over the degree to which the promulgated regulations mirror the intent of the original law, civil suits may be brought in court to resolve disputes. Judgments in such suits can and have had important effects on the CFR. In the remainder of this section we will explore briefly the body of air pollution law, from enabling legislation to promulgation of regulations in the CFR.

1. Federal Air Pollution Law

Basic federal law is contained in the United States Code (USC). It is divided into "Titles" which are themselves divided into "Sections." Groups of sections form "Chapters." Title 42 of the USC (usually denoted as 42 USC) is entitled "The Public Health and Welfare." It contains the basic law pertaining to air pollution: Chapter 15B entitled "Air Pollution Control" and Chapter 55 entitled "National Environmental Policy."

The Clean Air Act is contained in Section 1857 of Title 42 (within Chapter 15B) and is referenced by the notation 42 USC §1857. Originally enacted in 1963, it has since been amended a number of times. The most notable changes occurred with the passage of the Clean Air Act Amendments of 1970 and 1977, the former of which, among other things, created the Environmental Protection Agency (EPA), authorized the setting of national ambient air quality standards (NAAQS) and required the development of state implementation plans (SIPs) for the attainment of compliance with the NAAQS. After passage by the Congress and signature by the President, a bill containing such amendments or providing for new portions of the USC becomes a part of the public law and is referred to both by the Congressional session and a passage sequence number. The 1970 Amendments, for example, are referred to as Public Law 91-604. For reference, the 91st Congress convened for the two years from January 1969 to January 1971.

The other legislation most heavily affecting air pollution law is the National Environmental Policy Act (NEPA) of 1969 (Public Law 91-190), which amended Chapter 55 (National Environmental Policy) of Title 42. In its primary features, the act created the Council on Environmental Quality reporting to the President and mandated the preparation of environmental impact statements (EISs) for "major Federal actions significantly affecting the quality of the human environment." These are required for federal agency actions and for projects supported "in whole or in part" with federal financing. The NEPA is found in 42 USC §4321, 4331 to 4335, 4341, and 4341 to 4347.

2. The Code of Federal Regulations

Implementation of federal law is accomplished by promulgation of specific regulations, the body of which is contained in the Code of Federal Regulations. The CFR is divided into "Titles" (not the same as those in the USC), which are themselves subdivided into "Chapters," "Subchapters," and "Parts." All federal regulations pertaining to air pollution are contained in Title 40 which is called "Protection of the Environment." The formal organization of 40 CFR is shown in Exhibit III-1. Note that Title 40 contains no Chapters II and III.

Subchapter C, "Air Programs," is expanded in that exhibit to include "Part" subheadings as is Chapter V, "Council on Environmental Quality." The following parts within Chapter I are of particular importance. In Part 50 the primary and secondary NAAQS are set for sulfur dioxide, particulate matter, carbon monixide, photochemical oxidants, hydrocarbons, and nitrogen dioxide. In Part 51 requirements are stated for the development of SIPs. All State plans, whether approved or disapproved, are published in Part 52. In Part 60 the emissions standards are set for new and modified stationary sources. Further breakdown of these parts by section heading is provided in Appendix A.

As originally conceived, SIPs were blueprints for achieving compliance with the NAAQS. As the regulations have evolved, however, they now require that SIPs now provide for air quality maintenance (AQM) once compliance has been achieved. SIPs are currently being revised according to the mandates of the 7 August 1977 Clean Air Act Amendments and are required to be reassessed periodically as to their ability to attain and maintain the NAAQS.

EXHIBIT III-1. FORMAL ORGANIZATION OF CFR TITLE 40--PROTECTION OF ENVIRONMENT

Chapter 1. Environmental Protection Agency

- Subchapter A - General (Parts 0-21)
- Subchapter B - Grants and Other Federal Assistance (Parts 30-49)
- Subchapter C - Air Programs (Parts 50-89)
  - Part 50. National primary and secondary ambient air quality standards
  - Part 51. Requirements for preparation, adoption, and submittal of implementation plans
  - Part 52. Approval and promulgation of implementation plans
  - Part 53. Ambient air monitoring reference and equivalent methods
  - Part 54. Prior notice of citizen suits
  - Part 55. Energy related authority
  - Part 60. Standards of performance for new stationary sources
  - Part 61. National emission standards for hazardous air pollutants
  - Part 79. Registration of fuels and fuel additives
  - Part 80. Regulation of fuels and fuel additives
  - Part 81. Air quality control regions, criteria, and control techniques
  - Part 85. Control of air pollution from new motor vehicles and new motor vehicle engines
  - Part 86. Control of air pollution from new motor vehicles and new motor vehicle engines: certification and test procedures
  - Part 87. Control of air pollution from aircraft and aircraft engines
  - Part 88-89. [Reserved]
- Subchapter D - Water Programs (Parts 100-149)
- Subchapter E - Pesticide Programs (Parts 162-180)
- Subchapter F - [Reserved]
- Subchapter G - Noise Abatement Programs (Parts 201-210)
- Subchapter H - Ocean Dumping (Parts 220-230)
- Subchapter I - Solid Wastes (Parts 240-399)

Subchapter N - Effluent Guidelines and Standards (Parts 401-460)

Subchapter Q - Energy Policy (Part 600)

Chapter IV. Low Emissions Vehicle Certification Board (Part 1400)

Chapter V. Council on Environmental Quality (Parts 1500-1510)

Part 1500. Preparation of environmental impact statement: Guidelines

Part 1510: National oil and hazardous substances pollution contingency plan

Contained within SIPs are procedures for controling emissions from both mobile and stationary sources. Because of the size and age of the vehicle fleet, control of emissions from mobile sources is currently an important part of other SIP segments dealing with NAAQS compliance. As stricter automotive emissions standards are achieved and older cars are removed from highways through age attrition, stationary sources will contribute an increasing fraction of the total emissions inventory. Their importance thus increases in the AQM segment of the SIPs.

The portion of 40 CFR relating to the review of applications for new or modified stationary sources is Section 51.18. There it is stated that "no approval to construct or modify will be granted unless the applicant shows to the satisfaction of the Administrator that the source will not prevent or interfere with attainment or maintenance of any national standard." The quote is a paraphrase of §51.18(a), as written in the California SIP [40 CFR §52.233(g)(3)]. Several issues of practical importance derive from this section of 40 CFR. New source review (NSR) procedures are thus required, with such stationary sources directed to meet new source performance standards (NSPS) where stated in 40 CFR §60 or as determined by the appropriate reviewing agency and to install appropriate pollution control equipment. Also, an important consequence of 40 CFR §51.18 derives from its interpretation in urban areas currently in noncompliance with the NAAQS. In most instances, the addition of a single, modestly sized, stationary source would be unlikely to affect regional peak pollutant concentration. Considered separately, an argument could be made that few new stationary sources violate the letter of §51.18. Taken in the aggregate, however, emissions from several new sources together could have serious adverse effects on regional pollutant concentrations. To overcome this interpretive difficulty, the EPA has employed the so-called offset rules (OSR). All new stationary sources in noncompliant urban areas are considered to be in violation of §51.18 unless the applicant can demonstrate that a reduction in emissions from other sources and a reduction in the air quality impact of those emissions has been achieved to offset those produced by the proposed new source.

Another issue of importance in SIP development is the prevention of significant deterioration (PSD) of the air quality in areas currently in attainment of the NAAQS. Originally, 40 CFR contained no provision for consideration of PSD. A court suit, however, brought about a judgment that SIPs must address this issue. As a consequence, subsequent to May 31, 1972, the EPA Administrator disapproved all SIPs not considering PSD. Standards for PSD were promulgated in §52.21, entitled "Significant Deterioration of Air Quality."

In addition to SIPs, environmental impact statements and reports (EIS/R) represent the other major class of planning documents formally required to address air quality issues. In Chapter V of 40 CFR, guidelines are provided for drafting EISs for major federal actions. They are required not only for projects undertaken solely by the federal government, but also for any major projects supported "in whole or in part" by federal financing. EISs were submitted to the CEQ for review. They are now, however, received and reviewed by the EPA. State and local agencies can also require for individual projects a formal statement of environmental impact. In California, for instance, such a statement is called an "Environmental Impact Report" (EIR) and is filed pursuant to the California Environmental Quality Act (CEQA).

Running throughout air pollution law is the basic right of legal appeal. Court suits have played an important part in shaping the body of the law. Portions of the authorizing statutes, the CFR, and many individual EIS/Rs have come under legal challenge. As a result, litigation (LIT) also represents an important class of issues addressed by air pollution modelers.

B. GENERIC ISSUE CATEGORIES

In the previous section we have outlined many of the important features of air pollution law. A number of generic issues thereby have been identified. In this section we will summarize these generic issues, discuss each briefly, and then examine their implications for air pollution modeling.

In the next chapter we will match these issue categories with a number of existing models, comparing application requirements with model capabilities.

1. The Issues: Their Classification

The air pollution burden in a geographical area is the result of the complex interaction of emissions from all sources as they mix and disperse in the atmosphere, subject to prevailing influences of meteorology, solar irradiation, and terrain. The total pollutant concentrations experienced are a function of the effects of emissions from each of the mobile and stationary emitting sources, though that function is generally not a linearly additive one. Because the NAAQS are expressed in terms of total allowable concentration levels and are applicable at any location to which the public has access, implementation plans are inherently regional in perspective. There is a certain duality of focus in SIPs, however: While they detail plans for regional NAAQS compliance and maintenance, they do so through curtailment of emissions from individual sources and source categories. Thus, while the focus is ultimately on regional effects, the environmental impact of individual sources also must be considered. This is an explicit issue with new source review (NSR), for instance. As the number of sources to be considered decreases, the two perspectives--regional and single source-specific--merge together. A case in point is the examination of the impacts of a few sources located in a rural area, where prevention of significant deterioration (PSD) is an issue.

From the discussion of air pollution law presented earlier, we have isolated several specific issues, each falling into one of two distinct generic issue categories. The chief distinction between the two is not simply the difference between regional and source-specific perspective, for each individual source has both a regional and a localized downwind impact. Rather, the clearest distinction lies in the number of sources considered. Questions of regional NAAQS compliance and maintenance are multi-source issues. NSR, on the other hand, primarily concerns a single source. Using such a distinction, the principal issues addressed by air quality planners are as follows:

- Multiple-Source Issues
  - SIP/C (State Implementation Plan/Compliance). The attainment of regional compliance with the NAAQS, as considered in the SIP.
  - AQMP (Air Quality Maintenance Planning). Regional maintenance of compliance with the NAAQS, as considered in the SIP.
- Single-Source Issues
  - PSD (Prevention of Significant Deterioration). Limitation of the amount by which the air quality can be degraded in areas currently in attainment of the NAAQS; this is considered in each SIP.
  - NSR (New Source Review). Permit process by which applicants proposing new or modified stationary sources must demonstrate that both directly and indirectly caused emissions are within certain limits and that the pollution control to be employed is performed with the appropriate technology; this is considered in each SIP.
  - OSR (Offset Rules). Interpretive decision by which all new or modified stationary sources in urban areas currently in noncompliance with the NAAQS are judged unacceptable unless the applicant can demonstrate a plan for reducing emissions in existing sources and that a reduction in the air quality impact of these emissions has been achieved to offset those produced by the proposed new source; this decision has a strong impact on the stationary source permit process.
  - EIS/R (Environmental Impact Statement/Report). A statement of impact required for major projects undertaken by the federal government or financed by federal funds (EIS), or a report of project impact required by state or local statutes (EIR).
  - LIT (Litigation). Court suits brought to resolve disagreement over any of the issues mentioned above or to secure variances waiving federal, state or local requirements.

The above seven issues are classified according to their most frequently encountered form. We note that actual cases do not always conform to the bounds of the generic issue categories as shown. An EIS, for instance, can have a regional perspective, as with the Denver Overview EIS recently completed for Region VIII of the EPA. Also, LIT can occasionally have effects on regional NAAQS compliance and maintenance. For example, PSD and AQMP resulted from court suits.

2. The Issues: Some Practical Examples and Their Implications for Air Pollution Modeling

Many practical examples can be found in which the issues identified above play an important role in planning. At this point, we will discuss some of the more important applications in which they are likely to be encountered. Modeling requirements can thus be identified. This discussion will serve as a prelude to the examination of air pollution models presented in the next chapter.

First, we consider the nature of multiple-source (M/S) issue applications. SIP/C and AQMP can focus both on urban areas as well as on large rural sources. Here we concentrate on the most frequently encountered applications, those in urban areas. Encountered in such regions are both reactive pollutants [ozone (O_3), hydrocarbon (HC), and nitrogen dioxide (NO_2)] and relatively nonreactive pollutants [carbon monoxide (CO), sulfur dioxide (SO_2)*, and total suspended particulates (TSP)]. There are a variety of different source types: point sources (power plants, refineries, and large industrial plants, such as steel, chemical and manufacturing companies), line sources (highway, railroads, shipping lanes, and airport runways), and area sources (home heating, light industrial users of volatile chemicals, street sanding, gasoline distribution facilities, and shipping ports). Mobile sources (cars, trucks, and buses) almost invariably can be aggregated into highway line sources. While a few cities with air pollution problems are located in complex terrain (Pittsburg, for example), most are situated in relatively flat or gently rolling terrain. Geographical features can play an important part in regional air pollution (for instance, the ocean near Los Angeles, the lake near Chicago, and the mountains near Denver).

---

* Sulfur dioxide is slowly reactive: $SO_2 \rightarrow SO_4^=$, aerosol.

Air pollution modeling in such circumstances has been used for several principal purposes. It has been useful in estimating the total amount of emissions cutback required to reach compliance with the NAAQS. Individual control strategies also have been assessed, both for SIP/C and AQMP. Insights from regional modeling have been useful in modifying and improving pollutant measurement network design. In Denver, for instance, use of the SAI Urban Airshed Model indicated for a particular model day the presence of an ozone (O_3) peak in a then-unmonitored area. Subsequent location of a temporary monitoring station at that site lead to the observation of O_3 readings in excess of any previously measured. Also, models have had an influence on transportation network design (the balance of freeways, arterials, and feeders) and modal split (the mix between personal and mass transit). Through the EIS/R process, individual projects (for example, the Interstate 470 freeway and the construction of wastewater treatment facilities, both in Denver) have been examined using models to estimate air quality impact.

Second, we consider the nature of stationary single-source (S/S) issues. Important applications occur in both urban and rural areas. These focus on the following: (1) SIP/C and the permit approval process for new or modified stationary sources and (2) the variance process for existing facilities. As for the first of these, SIP/C and the permit approval process, all new or modified major S/Ss, urban and rural, are subject to NSR and must meet NSPS and use the best available pollution control equipment. Also, both direct and indirect impact on air quality must be considered.

In urban areas, major S/Ss might include proposed refineries, power plants, and industrial facilities, as well as shopping, employment, and recreational/sports centers. With the last of these, indirect effects are particularly important. Each draws appreciable numbers of automobiles, adding to local vehicle miles traveled (VMT) and increasing congestion and thus pollutant emissions. Also, automobile hot soak and some cold start emissions are concentrated in accompanying parking lots.

Urban S/Ss are dealt with in the SIP/C and permit application process differently than are rural S/Ss. In urban areas in noncompliance with the NAAQS, OSR must be considered. The air pollution modeler must be able not only to represent the regional and localized downwind impact of the new S/S but also to estimate the subtractive effect of reducing emissions from one or more existing sources.

Another difference between urban and rural areas has important significance for the modeler. In rural areas, the relatively nonreactive pollutants (SO_2 and TSP) are often of greater interest than are the more reactive ones. Although the NO_x emissions also produced at some point could generate, with the addition of HC, photochemically reactive pollutants, they are usually not of primary concern. In urban areas, the reactive pollutants (O_x, NO_2, and HC) must also be modeled. When the incremental effect of a S/S is being considered in an urban areas (OSR, as well), this distinction can have a strong effect on model choice. This is particularly true when an S/S emits O_x precursors such as NO_x, which power plants do, or HC, which refineries do.

In rural areas, applications centering on energy development have been prominent in recent years, particularly in the northern and central Great Plains. The direct air pollution impact of these S/Ss would be produced by coal extraction (strip mining), conversion to natural gas, transport to energy production facilities if they are not on site (via unit train or slurry pipelines), or coal combustion in large power plants. Indirect impact would result from the construction of the above-mentioned facilities (new highways, provision for temporary construction crews) and the growth of nearby "boom" towns (housing for families of workers and the additional population increase required to provide commercial and public services to workers).

A complicating factor not confronted in nonattainment regions arises in attainment areas: PSD must be considered. No S/S or combination of them is permitted to degrade significantly the air quality in nonpolluted rural areas. In each SIP such areas are identified. The modeler must be able to assess the likelihood that an S/S will impinge on such areas to an unacceptable degree. Also, because pollutants from rural sources are either inert or slow in reacting and because surface deposition, rainout, and washout often proceed at slow rates (depending on synoptic meteorology), atmospheric residence times are long for some pollutants such as the derivative products of SO_2. Transport distances on the order of a thousand kilometers may not be unusual. The modeler must be able to account for pollutant transport and transformation on this temporal and spatial scale, if required.

In both urban and rural areas, the owner of a S/S has the right to seek a variance temporarily excusing the source from provisions of the law, but not such as to cause a violation of the NAAQS. A number of reasons could motivate such a request. For a power plant, petroleum shortages could result in a need to burn high-sulfur fuel. For a refinery, petroleum storage and shipping needs might result in a variance request. Other reasons might include a need for an extension of the time required to comply with SIP control strategy requirements or for periodic pollution control equipment maintenance or replacement.

3. The Issues: A Prologue to the Next Chapter

In this chapter, we have examined the body of air pollution law and identified two generic issue categories: multiple-source issues and single-source issues. Seven separate (though interrelated) types of issues were classified within that structure: SIP/C, AQMP, PSD, NSR, OSR, EIS/R, and LIT.

We have examined some practical examples illustrating particular features of these issues as they manifest themselves in both urban and rural areas. We have also discussed some key implications that these issues have for air pollution modeling. This serves as an important prologue to the discussion of specific models undertaken in the next chapter. In that chapter we will match application requirements to model capabilities. The issues identified here will serve as the framework within which that discussion is carried out.

IV AIR QUALITY MODELS

In the last chapter, we identified generic types of air quality issues. In this chapter, we define generic classes of models. Having done so, we match the two, identifying those issues for which each model may be a suitable analytical tool. We also describe the technical formulations and underlying assumptions employed in each generic model class, indicating some key limitations.

The final choice of a model for use in addressing a particular issue can be made only by considering the characteristics of the proposed application. To facilitate the comparison between model capabilities and applications requirements, we define a set of applications attributes. We then match the two, identifying for each generic model the combinations of application attributes for which it is suited. A related means for matching model to application is described in EPA (1978a).

In this chapter we attempt to specify the relationship between issues, models, and applications. Having done so, we then develop in Chapter V model performance measures appropriate to each issue/model combination of practical interest. This will set the stage for a discussion of requisite model performance standards in Chapter VI.

In order to preserve generality, our emphasis in this chapter centers primarily on generic model categories rather than on specific air quality models. Certain benefits may be achieved thereby: General conclusions appropriate to an entire class of models may be stated without reference to any particular model, and extensive discussions of any observed differences between intended capabilities and technically achieved ones need not be conducted for each specific model.

Our central purpose in this report is to discuss means for setting model performance standards. While not central to this, however, we do recognize a need to associate some specific models with our generic model categories. To assist in doing so, we examine in Appendix B a number of air quality models. Though the list is not a complete one, a number of available models are examined in detail and tabulated according to several attributes. Among these are the following: level of intended usage (screening or refined), type of pollutant (reactivity, averaging time), degree of resolution (spatial and temporal), and certain site specifics (terrain, geography, as well as source type and geometry).

We summarize at the end of this chapter that part of Appendix B needed to associate specific models with our generic categories. No attempt is made in this chapter or in Appendix B to screen models for technical acceptability nor is any attempt made to be all-inclusive. Models are classified according to their intended capabilities rather than their technically achieved ones. Among the references we have drawn upon in gathering this information are the following: Argonne (1977), EPA (1978b), and Roth et al., (1976), as well as several program users' manuals.

A. GENERIC MODEL CATEGORIES

In this chapter air quality models and prediction methods are classified into generic model categories. Here we describe the structure of the classification scheme employed, the full form of which is shown in Exhibit IV-1. Though many such schemes have been proposed (Roth et al., 1976, and Rosen, 1977, for example), we identify three broad divisions: rollback, isopleth, and physico-chemical. We describe here each of these categories, mentioning technical formulation, general capabilities, and major limitations. In doing so, we draw upon material in Roth et al. (1976).

1. Rollback Category

Included in the first of these are all those prediction methods in which ambient pollutant concentrations are assumed to be directly (though not necessarily linearly) proportional to emissions, according to some simple relationship. Emissions control requirements are presumed proportional to the amount by which the peak pollutant concentration exceeds the NAAQS. Linear rollback and Appendix J are examples of such methods.

- I. Rollback
- II. Isopleth
- III. Physico-Chemical
  - A. Grid
    - 1. Region Oriented
    - 2. Specific Source Oriented
  - B. Trajectory
    - 1. Region Oriented
    - 2. Specific Source Oriented
  - C. Gaussian
    - 1. Long-Term Averaging
    - 2. Short-Term Averaging
  - D. Box

EXHIBIT IV-1. GENERAL MODEL CATEGORIES

Because atmospheric processes are generally complex and nonlinear, the fundamental proportionality assumption invoked in rollback methods is frequently violated in actual application. For this reason, rollback methods are usually regarded as screening techniques, whose results give at best only a general indication of the amount of emissions control required. They are most often used when insufficient data are available to perform an analysis that is more technically justifiable. Even then, results obtained with them are appropriate only as a crude indication of the need for more extensive data gathering and analysis. Because rollback methods lack spatial resolution, they are most suitable for addressing regional, multiple-source* issues. Also, their use is more appropriate for applications involving relatively nonreactive pollutants (SO_2, CO and TSP).

---

* In this report, "multiple-source" refers to many, well-distributed sources of all types and sizes. It does not include, for instance, a single complex having multiple stacks.

2. Isopleth Category

Within the second generic model category are included those methods relying on isopleth diagrams to relate precursor concentrations of primary emissions (usually oxides of nitrogen and nonmethane hydrocarbon) to the level of secondary pollutant (usually ozone) resulting from such a mixture. As is true with the EPA EKMA method (see EPA, 1977), these diagrams are usually constructed from computer simulations using theoretically and chamber derived chemical kinetic mechanisms. They invoke assumptions about a number of parameters such as regional ventilation and solar insolation, as well as pollutant entrainment, carryover from the previous day, and transport from upwind. The accuracy of the postulated chain of chemical reactions is evaluated using smog-chamber data. The types of information required to construct an isopleth diagram are roughly equivalent to those required to employ a box model, and we note that the two methods are conceptually similar in many regards. We maintain a distinction between the two, however, because of the view prevailing in the user community that they are separate classes of models. Also, not all box models are photochemical, as are isopleth-based methods.

Entry into an isopleth diagram requires an estimate of the peak concentration actually occurring during the day on some initial base date. Given an assumption about the relative proportion of precursor species control (HC versus NO_x), the degree of emissions cutback required to achieve the NAAQS can be estimated directly.

Isopleth methods lack spatial resolution. They are thus capable of addressing only regional, multiple-source issues. By their nature, isopleth methods are useful only for applications involving photochemically reactive pollutants. Because of the level of approximation involved in constructing the isopleth diagram itself, in entering it using measured ambient data, and in accounting for the effect of transport from upwind, such methods are more appropriate for use as screening tools. In this capacity, they can be helpful in assessing the need for further, more refined analysis. However, in some limited applications where the assumptions invoked in the formulation of the isopleth methods are generally satisfied, estimates of the required degree of emissions control obtained using such a method can be regarded as acceptably accurate.

3. Physico-Chemical Category

The third category contains models based upon physical and chemical principles as embodied in the atmospheric equations of state. It is divided into four main subcategories: grid, trajectory, Gaussian, and box. We discuss here each subcategory.

a. Grid Subcategory

Grid models employ a fixed Cartesian reference system within which to describe atmospheric dynamics. The region to be modeled is bounded on the bottom by the ground, on the top usually by the inversion base (or some other maximum height), and on the sides by the desired east-west and north-south boundaries. This space is then subdivided into a two- or three-dimensional array of grid cells. Horizontal dimensions of each cell measure on the order of several kilometers, while vertical dimensions can vary, depending on the number of vertical layers and the spatially and temporally varying inversion base height. Some grid models assume only a single, well-mixed cell extending from the ground to the inversion base; others subdivide the modeled region into a number of vertical layers.

Ideally, the coupled atmospheric equations of state, expressing conservation of mass, momentum, and energy, would be solved systematically within each grid cell, with a chemical kinetic mechansim used to describe the evolution of pollutant species. Several major difficulties arise in practice. Computing limitations are rapidly encountered. A region fifty kilometers on a side and subdivided into five vertical layers requires 12,500 separate grid cells if grid cells are one kilometer on a side. Maintaining a sufficient number of species to allow the functioning of a chemical kinetic mechanism compounds the storage problem. For a ten-species mechanism, storage of the concentrations for each species in each grid cell in our example would alone require 125,000 storage locations.

To avoid these and other computing or numerical problems, most grid models solve only one atmospheric state equation--the conservation of mass, or continuity, equation, decoupling the other two. The momentum equation is replaced by meteorological data supplied to the model in the form of spatially and temporally varying wind fields. The energy equation is supplanted by externally supplied vertical temperature profile data, from which inversion heights are also calculated.

Other problems are encountered in solving the mass continuity equation, a principal such problem being the atmospheric viscosity terms. Turbulence, which is a randomly varying quantity, can be described only in statistical terms. Species concentrations, as a result, can be found only as values averaged over some time interval. Also, the continuity equation can be solved only if turbulence effects are decoupled through a series of approximations involving turbulence gust eddy sizes and strengths.

Grid models require the specification of time-varying boundary conditions on the outer sides and the top of the modeled region, the initial conditions (species concentrations) in each grid cell at the start of a simulation, and spatially and temporally varying emissions for each primary pollutant species. The first two of these are derived from station measurement data, and the last is obtained from an appropriate emissions inventory for the modeled region.

Grid models are capable of considering both reactive and relatively nonreactive pollutant species. Models considering reactive species, because of their limited time scale (less than several days), are appropriate tools only for addressing questions involving pollutants having short-term standards (O_3, CO, HC, and SO_2) and for medium-range pollutant transport (an urban plume, for example). Some grid models are designed to model large spatial regions (such as the Northern Great Plains--see Liu and Durran, 1977) and thus can address long-range transport questions. At their present state of development, these models are appropriate tools only for examining questions involving relatively nonreactive pollutants (principally long-term SO_2 and TSP).

There are two major classes among grid models: region oriented and specific source oriented. In the first class, two basic variants exist: urban scale and regional scale models. The first of these attempts to model the urban environment, considering emissions from a number of different sources and simulating both reactive and relatively non-reactive pollutant species over a spatial scale on the order of tens of kilometers through a temporal scale of 8 to 36 hours. Regional-scale models, on the other hand, represent an attempt to model long-range pollutant transport over a spatial scale of hundreds of kilometers through a temporal scale of several days. Emissions are assumed to come from a few widely dispersed, usually rural, sources; the pollutants considered are relatively nonreactive (or more precisely, slowly reactive) ones such as SO_2. (Though $SO_2 \rightarrow SO_4^=$, it does so much more slowly than the time scale of reactions involving the more reactive species.) One such model was developed by SAI for use in assessing the air quality impact of large-scale energy development in the Northern Great Plains (Liu and Durran, 1977).

Because of their spatial extent, regional oriented grid models are appropriate tools for addressing regional (multiple-source) issues, such as SIP/C and AQMP. Because of their spatial resolution, certain regional questions about single-source issues can also be addressed. The regional effect of a new source can be assessed. The subtractive regional effect of removing an existing source also can be estimated, an essential capability for addressing OSR questions. However, only grid models specifically designed to consider a single source have sufficient spatial resolution to assess near-source, or microscale effects.

Specific source oriented models represent the second major class of grid models. Some specific examples of such models are listed later in this chapter. Models of this type are particularly useful in two types of applications: examining the behavior of a plume containing reactive constituents, and accounting for the effects of complex terrain on a point source plume. Because of their formulation, these models can consider the effects of plume interaction with ambient reactive pollutants. This is of interest in addressing single-source issues in urban areas with significant levels of reactive pollutants. Often urban-scale grid models are used to predict the ambient conditions with which the plume interacts.

Those models designed for applications in complex terrain can be used when it is necessary to describe explicitly the wind fields and inversion characteristics encountered by a dispersing pollutant. Although simpler models exists, they are often inadequate when applied in situations in which terrain is particularly complex or when photochemical reactivity is important.

b. Trajectory Subcategory

Trajectory models employ a reference coordinate system that is allowed to move with the particular air parcel of interest. A hypothetical column of air is defined, bounded on the bottom by the ground and on the top by the inversion base (if one exists), which varies with time. Given a specified starting point, the column moves under the influence of prevailing winds. As it does so, it passes over emissions sources, which inject primary pollutant species into the column. Chemical reactions are simulated in the column, driven by a photochemical kinetic mechansim. Some trajectory models allow the column to be partitioned vertically into several layers, or cells. Emissions in such models undergo vertical mixing upward from lower cells. Other trajectory models allow only a single layer; in these, vertical mixing is assumed to be uniform and instantaneous.

The forumlation employed by trajectory models to describe atmospheric dynamics represents an attempt to solve the mass continuity equation in a moving coordinate system. The remaining state equations--conservation of momentum and energy--are not solved explicitly. As is done in grid models, solution of the momentum equation is avoided by specification of a spatially and temporally varying wind field, while solution of the energy equation is sidestepped by externally supplying temperature and inversion base height information.

Several basic assumptions are invoked in the formulation of trajectory models. Since only a single air column is considered, the effects of neighboring air parcels cannot be included. For this reason, horizontal diffusion of pollutants into the column along its sides must be neglected. This may not seriously impair model results so long as sources are sufficiently well distributed that emissions can be idealized as uniform, or nearly so, over the region of interest. However, if the space-time track of the air column passes near but not over large emissions sources, neglect of the effect of the horizontally diffusing material from those sources might cause model results to be deficient. In general, problems occur whenever there are significant concentration gradients perpendicular to the trajectory path.

Also, the column is assumed to retain its vertical shape as it is advected by prevailing winds. This requires that actual winds be idealized by means of a mean wind velocity assumed constant with height. Because of the earth's rotation and frictional effects at ground level, winds aloft usually blow at greater speeds than do surface winds, and in different directions. This produces an effect known as wind shear, which is neglected in trajectory models. If emissions are evenly distributed in amount and type over the region of interest and winds are also uniform, this may not represent a serious deficiency. In such a case, material blown out of the column by wind shear effects would be replaced by similar material blown into it, with the net effect on model results expected to be small. However, if a significant fraction of the emissions inventory is contributed by large point sources or if wind patterns display significant spatial variation, neglect of wind shear can seriously impair the reliability of trajectory model results.

Additionally, many trajectory models assume that the horizontal dimensions of the air column remain constant and unaffected by convergence and divergence of the wind field. Where winds are relatively uniform, this may not be of serious consequence. Where winds have significant spatial variation, as could be the case in even mildly complex terrain, however, this assumption could lead to deficient results. In the San Francisco Bay region, for example, wind flow convergence during the day causes the merging of several air parcels. Peak pollutant concentrations subsequently occur in this merged "super-parcel." A trajectory model would be an inadequate tool for addressing problems in such a region.

In general, trajectory models require as inputs much the same types of data required to exercise a grid model. Emissions are required along the space-time track of the air column. Wind speed and direction must be provided to determine its movement. Vertical temperature soundings must also be input in order to determine the height of the column (the height of the inversion base). Although these data need be prepared only for the corridor encompassing the trajectory path, general application of the model to an entire urban area requires that data be prepared for a significant portion of the region.

Two major classes exist among trajectory models: region oriented and specific source oriented. The first of these classes includes those models designed to address multiple-source, regional issues, usually in urban areas. The second class contains so-called reactive plume models. For reasons noted above, the use of trajectory models is appropriate on an urban scale only in certain circumstances. Careful screening is required of the emission and meteorological characteristics in a proposed application region to insure the appropriateness of trajectory model usage.

The second class of trajectory models includes those designed to evaluate the air quality impact downwind of a specific source. Because of the underlying equation formation, these models are more appropriate for use in areas having relatively simple terrain. However, because they are capable of simulating photochemical reaction, they can be used in addressing issues involving reactive pollutants. Often, region oriented models are used to generate the ambient conditions with which the reactive plume downwind of the source must interact. For all trajectory models considering reactive pollutants, the time scales remain short (less than several days). Consequently, they are inappropriate for consideration of problems involving pollutants subject to long-term standards.

c. Gaussian Subcategory

In the formulation of Gaussian models, the atmosphere is assumed to consist of many diffusing pollutant "puffs," all moving on individual trajectories determined by prevailing winds. The concentration at any point is assumed due to the superimposed effect of all puffs passing over the point at the time of observation. Rather than keeping track of the path of each puff, their motion (both advection and diffusion) is described in terms of conditional state transition probabilities. Given an initial location at a particular time, this state transition probability describes the likelihood that the puff will arrive at another specified point a given time interval later. With an entire field specified at some reference time, the net expected effect at a particular point and time is calculated by determining the integral sum of the separate expected effects of each puff in the field.

Central to this type of formulation is a knowledge of the time-varying state transition probabilities for the entire concentration field. In practice, turbulence nonuniformities and terrain-specific effects combine to render it unlikely that such probabilities can be determined. To overcome this difficulty, traditional Gaussian models (among others, those recommended by the EPA) invoke several assumptions. First, the turbulence field is assumed to be stationary and homogeneous, which implies it has two important qualities: First the statistics of the state transition probabilities can be assumed dependent only on spatial displacement, thus removing their time-dependency; and second, the probabilities are not dependent on puff location in the field, thus removing spatial variability. These are satisfactory approximations so long as significant differences do not exist between turbulence characteristics of the atmosphere in different portions of the region to be modeled. For applications in complex terrain, for instance, such an assumption might not be justified.

Once turbulence field stationarity and homogenity have been assumed, it still remains to specify the functional form of the state transition probability. Gaussian models derive their name from their assumption that this probability function is Gaussian in form. Given this assumption, the concentration field can be determined analytically by evaluating the integral expressing the summation of separate effects from all pollutant puffs affecting the region of interest. In order to isolate the effect of an individual source, only puffs containing pollutants emitted from that source are considered.

Concentrations about the plume centerline are assumed to be distributed according to a Gaussian relationship, whose vertical and horizontal cross-sectional shape is a function of downwind distance from the source and atmospheric stability class. Analytic forms can be determined expressing the form of the downwind concentration field for several different types of emissions regimes: instantaneous "puff," continuous point source emission (steady-state), continuous emissions from an area source, and continuous emissions along a line source.

Several other assumptions are invoked in Gaussian steady-state models. The vertical and horizontal spread of the plume is assumed characterized by dispersion coefficients, whose values are dependent on the distance downwind of the source. They are assumed to be functions of atmospheric stability and are thus characterized by stability class. Specific values are obtained from standard workbooks, such as that developed by Turner, or evaluation of data measured downwind of actual sources.

In many models, plume interaction with the ground and the inversion is considered. Usually, perfect or near-perfect reflection is assumed to occur. Multiple reflections are often modeled, although some models assume that beyond a certain downwind distance mixing is uniform between the ground and the inversion base.

Consideration of plume rise is made in Gaussian point source models. Depending upon ambient atmospheric conditons, such as temperature and humidity, hot gases from an emitting stack may rise, sink or remain at the same height. Simplifying thermodynamic equilibrium relationships, such as that developed by Briggs, are often used to estimate the magnitude of plume rise.

Two major classes of Gaussian models exist: long-term averaging and short-term averaging. Though both invoke the basic Gaussian assumptions, major differences exist in formulation. Long-term models divide the region surrounding each source into azimuthal sectors. The long-term variation of the wind at the source must then be specified by wind speed and direction (by sector) classes, along with the frequency of occurrence for each combination. This information usually is conveyed in the form of a "wind rose." Data describing the frequency of occurrence of the various atmospheric stability categories must also be specified. The probability of occurrence of of stability category/wind vector (speed and direction) combination is then used to weight the downwind concentrations resulting from it. The weighted sum represents the expected value of the long-term averaged pollutant concentration. Models employing this so-called "climatological" formulation are appropriate tools for addressing problems involving pollutants for which long-term (annual) standards are specified (SO_2, TSP, and NO_2).

The second class of Gaussian models includes those designed for short-term analysis. Prevailing wind direction and speed, as well as emissions characteristics, are assumed to persist long enough that steady-state conditions are established. The downwind concentration field resulting from source emissions can then be evaluated analytically. Some models allow a limited form of temporal variability by dividing the modeling day into segments (perhaps one hour long), during each of which conditions are assumed to be in steady state. Source strengths and prevailing wind speed at the height of emissions release are required for each segment, as are sufficient vertical temperature profile data to calculate inversion base height, if one exists, and atmospheric stability class. The last of these is required in order to determine vertical and horizontal dispersion coefficients. Because wind data frequently are not available at the height of emission release, surface wind measurements are extrapolated. Wind speed is assumed to vary vertically according to a power law, the exponent of which is given as a function of stability class. Determination of stability class is made by one of several appropriate methods, each of which is also dependent on surface observations.

Both Gaussian classes contain models that can be used to estimate the impact of single or multiple sources. Some models are designed to consider only a single point source; others can model many different sources simultaneously. Consequently, the first group of these is appropriate only for addressing single-source issues; the second group can be used to consider multiple-source issues as well. Most models in this second group, though able to account for many sources, can also simulate as few as one. They can thus be used to consider both single and multiple source issues.

Full consideration of regional-scale issues (SIP/C and AQMP) requires of a model the ability to simulate all types of sources: point, area, and line. Not all multiple-source Gaussian models are capable of doing so.

Some are used to consider only point and area sources; others are used to consider line sources only. These latter are usually intended for use in addressing traffic related questions; they might be used, for instance, to estimate the impact of emissions from a full highway network on regional CO distribution and level. Consequent to the above, consideration of all source types in a region may require the joint use of more than one model--one considering point and area sources and another simulating line sources.

An important restriction exists on the type of pollutant species that can be simulated using Gaussian models. Because the formulation cannot accommodate explicit kinetic mechanisms, only relatively nonreactive pollutants can be modeled (CO, TSP, and SO_2).* However, some models incorporate first-order, exponential decay to account for pollutant removal processes and limited species chemical conversion. Multiple-source Gaussian models assume that the combined effect of many emitters can be calculated by linearly superimposing the effects from each individual source. Such an assumption would be an erroneous one if questions involving reactive species were being considered.

Some Gaussian models have been designed to simulate the effects of point source emissions in complex terrain. Various assumptions are made about the behavior of the plume and the variation in height of the inversion base as an obstacle is approached. Usually the plume is allowed to impinge on the obstacle without any sophisticated means to account for flow alteration, although some models allow for flow convergence and divergence in the wind field. Also, the base of the inversion is sometimes assumed to be at constant height above the source; in other models it is assumed to be a fixed distance above the terrain, thus varying with it. However, the Gaussian formulation depends on the assumption of turbulence field stationarity and homogeneity. This is a simplification that may not be justified in many applications in complex terrain.

---

* Long-term Gaussian models are also used to model annual NO_2, a reactive species, for which no short-term standard currently is set. This usually is accomplished by combining NO and NO_2 as NO_x, the "species" modeled. NO_x exhibits less variability during the day than NO_2 taken separately.

d. Box Subcategory

Box models are the simplest of the physico-chemical models. The region to be modeled is treated as a single cell or box, bounded by the ground on the bottom, the inversion base on the top, and the east-west and north-south boundaries on the sides. The box may enclose an area on the order of several hundred square kilometers. Primary pollutants are emitted into the box by the various sources located within the modeled region, undergoing uniform and instantaneous mixing. Concentrations of secondary pollutants are calculated through the use of a chemical kinetic mechanism. The ventilation characteristics of the modeled region are represented, though only grossly, by specification of a characteristic wind speed.

Because of their formulation, box models can predict, at best, only the temporal variation of the average regional concentration for each pollutant species. Consequently, they are capable of addressing only multiple source, regional issues. Furthermore, such models are useful only in regions having relatively uniform emissions. In those areas where point sources contribute significantly to the emissions inventory (in number and amount), the assumption of emissions uniformity may be an unsatisfactory one.

Box models require only limited data. Emissions can be specified on a regional basis, eliminating any need for determining their spatial variation. Only simple meteorological data need be supplied as input. For these reasons, box models can be used when little information is available. They are more appropriately used as screening tools, helping to identify those situations requiring more extensive data collection and modeling analysis.

B. GENERIC ISSUE/MODEL COMBINATIONS

The discussion in the previous section outlined the characteristics of generic classes of air quality models. In this section we associate generic model type with generic issue category. In so doing, we indicate the gross suitability of a generic model type as a tool in addressing a particular issue. As noted earlier, each generic model (GM) has associated with it a set of limitations on its use. In Section C we summarize the effects of these limitations. We first classify types of actual applications according to several key attributes and then indicate those which each GM is capable of considering. The result is an enumeration of possible model/application combinations.

In order to match model to issue, we present in Table IV-1 a matrix of model/issue combinations. For each GM, an indication is provided of its usefulness in addressing each of the seven generic issues identified in the previous chapter. Even where a GM is indicated as suitable, however, its inherent limitations (some of which are noted in the table) may prevent its use in certain applications. Consequently, further examination is required in order to make a final GM selection.

Summarizing the basic features of Table IV-1, we note the following:

> Grid Models

- Region Oriented Models. Urban scale models are able to address multiple-source issues (SIP/C, AQMP) involving both reactive and nonreactive pollutants. Their short-term temporal scale (< 36 hours), however, restricts them to problems involving pollutants with short-term standards (O_3, HC, CO, and secondary SO_2). Their spatial resolution (on the order of tens of kilometers) allows them to address some single-source issues (OSR, EIR, LIT). Regional scale models, as opposed to urban scale ones, are more oriented towards application in rural areas (few sources) involving nonreactive (or rather, slowly reactive) pollutants, such as SO_2, TSP, CO, and NO_2, which is slowly reactive in nonurban areas because of limited ambient HC). Their short-term temporal scale (on the order of a week or less), often a practical restriction due to computing requirements, limits their use in predicting long-term pollutant concentrations (SO_2, TSP, NO_2). They are suited for addressing questions involving single-source issues (PSD, NSR, EIS/R, LIT) in isolated rural areas.

TABLE IV-1. AIR QUALITY ISSUES COMMONLY ADDRESSED BY GENERIC MODEL TYPE

	Issue Category						
Generic Model Type	STP/C	AQMP	PSD	NSR	OSR	EIS/R	LIT
Refined Usage							
1. Grid[1]							
a. Region Oriented	X	X	X	X^2	X	X	X
b. Specific Source Oriented			X	X	X^3	X	X
2. Trajectory[1]							
a. Region Oriented	X	X			X	X	X
b. Specific Source Oriented			X	X	X^3	X	X
3. Gaussian[3]							
a. Short-Term Averaging[1]							
i) Multiple Source	X	X	X		X	X	X
ii) Single Source	X		X	X	X	X	X
b. Long Term Averaging[4]	X	X	X	X	X	X	X
Refined/Screening Usage							
4. Isopleth[1,5]	X	X					
Screening Usage							
5. Rollback	X	X					
6. Box	X	X					

Notes:

1. Only short-term time scales can be considered (less than several days).
2. Regional impact of new sources can be assessed but not near-source, or microscale, effects.
3. Only non-reactive pollutants can be considered.
4. Only pollutants having long-term standards can be considered (SO_2, TSP, and NO_2).
5. Only photochemically active pollutants can be considered.

    - Specific Source Oriented Models. These models are used primarily for addressing single-source issues (PSD, NSR, OSR, EIS/R, LIT). This class contains the so-called reactive plume models. Their ability to consider reactive pollutants makes them suitable for urban applications or rural applications where plume reactivity is important. However, because OSR (a primarily urban issue) requires an estimate of the subtractive effect of removing an existing source, only questions involving pollutants for which linear superposition is approximately valid, i.e., nonreactive pollutants, can be addressed in an urban area with a specific-source model. These models are also suitable for use in applications where terrain complexity is important.

> Trajectory Models

    - Region Oriented Models. With some important restrictions, these models can be suitable for use in addressing multiple-source issues (SIP/C and AQMP) and, in limited circumstances, some single-source issues (OSR, EIS/R, LIT). Among the most important of such restrictions are the following: Emissions must be approximately uniform over the modeling region; air flow cannot be complex enough to cause merging of air parcels, i.e., flow convergence or divergence should not be important; and horizontal diffusion effects should not have significant nonuniformities, e.g., large point sources near but not within the space-time track of the advected air parcel being modeled. Because chemical kinetic mechansims can be included in their formulation, these models are capable of considering reactive as well as nonreactive species. Their temporal scale is so short, however, that no estimates of long-term concentration averages can be computed.
    - Specific Source Oriented Models. Subject to the same restrictions mentioned above, these models can be appropriate tools for use in considering single-source issues (PSD, NSR, OSR, EIS/R, LIT). Because they can consider reactive pollutant

species, they can be used in applications involving reactive plumes. Limited terrain complexity can also be simulated, so long as the abovementioned restrictions are not violated.

> Gaussian Models

Long-Term Averaging Models. These models can be used to address both multiple-source issues (SIP/C, AQMP) and some single-source issues (PSD, OSR, EIS/R, LIT). Because of the Gaussian formulation they cannot consider chemistry or surface removal effects beyond first order, i.e., exponential decay. Thus, they are appropriate tools only for addressing questions involving nonreactive (slowly reactive) pollutants. Their temporal scale is such that only pollutants having long-term (annual) standards can be considered (SO_2 primary standard, TSP, NO_2, where NO_2 is taken as NO + NO_2, i.e., NO_x). As currently configured, these models are appropriate for use in both urban and rural settings, although the terrain in such applications should be relatively simple.

- Short-Term Averaging Models. Two variants exist among these models: multiple-source and single-source. The types of issues they may be used to address divide similarly. Some multiple-source models, however, do not consider all types of sources: Some consider only point and area sources; others consider only line sources. The latter group is useful for examining the effects of traffic-related pollutants (particularly CO) resulting from highway network emissions. Consequently, if regional questions are to be addressed, the concurrent use of more than one model may be required. Only relatively nonreactive pollutants may be examined using this type of model. Because of their short-term temporal scale, these models are best suited for addressing questions involving pollutants having short-term standards (CO, SO_2 secondary standard).

> Rollback Models

Because rollback models lack spatial resolution, they are appropriate only for considering questions involving multiple-source issues (SIP/C, AQMP). Their use is generally confined to urban areas located in simple terrain. Their assumption that emissions are directly proportional to peak pollutant values is a technically limiting one. Consequently, they should be viewed as screening tools to evaluate the need for more extensive analysis and data gathering.

> Isopleth Models

Lacking spatial resolution, isopleth models are appropriate only for use in addressing multiple-source issues (SIP/C, AQMP). Employing ozone isopleth diagrams derived through the use of a photochemical kinetic mechansim, these models are designed to examine questions involving reactive pollutants (O_3, HC, short-term NO_2). Their use is most appropriate for applications in urban areas located in simple terrain. Because the isopleth diagram is constructed using regional ventilation, emissions, and background/transport assumptions, it is similar to the box models, which are described below. Like the box model, its technical limitations, except under exceptional circumstances, render it more useful and reliable as a screening tool to evaluate the need for more extensive analysis.

> Box Models

Because they lack spatial resolution, box models are appropriate only for use in considering multiple-source issues (SIP/C, AQMP). They assume spatially uniform emissions. For this reason, their use is more suited to areas that are urban or semi-urban. They are best used in modeling areas lcoated in simple terrain but have also been used in applications in complex terrain. An example of the latter type of application might be the modeling of a mountain valley containing several ski resorts and related developments. Technical limitations render the box models more suitable as screening tools.

C. MODEL/APPLICATION COMBINATIONS

In the previous section we discussed the relationship between generic models and generic issues. In this section we associate those generic models and the specific applications in which they may be used. We first classify applications by means of several key attributes. We then compare the possible values of these with model capabilities. For each generic model type, we are thereby able to identify the range of applications for which the model is suited.

Applications are characterized here by five attributes: number of sources, area type, pollutant, terrain complexity, and required resolution. In Table IV-2 we list the possible designations these attributes may assume. Against these we match generic model capabilities, identifying the list of designations for which each is suitable. A chart of the resulting model/application combinations is presented in Table IV-3. While exceptions may occur, the list of attribute designations shown is chosen based upon considerations presented earlier in this chapter.

D. SOME SPECIFIC AIR QUALITY MODELS

Our central purpose in this report is to discuss means for setting suitable standards for model performance. As prologue to this, both air quality issues and the models used to address them needed to be examined. We have done so in general terms to this point. Throughout this discussion we have referred to air quality models only in generic terms. By doing so, several advantages were achieved: General conclusions appropriate to an entire class of models could be stated without reference to any specific model, and extensive discussions of any observed differences between intended capabilities and technically achieved ones were not necessary for each particular model.

TABLE IV-2. POSSIBLE DESIGNATIONS OF APPLICATION ATTRIBUTES

Attribute	Possible Designations
Number of Sources	Multiple-Source Single-Source
Area Type	Urban Rural
Pollutant	Ozone (O_3) Hydrocarbon (HC) Nitrogen Dioxide (NO_2) Sulfur Dioxide (SO_2) Carbon Monoxide (CO) Total Suspended Particulates (TSP)
Terrain Complexity	Simple Complex
Required Resolution	Temporal Spatial

TABLE IV-3. MODEL/APPLICATION COMBINATIONS

Generic Model Type	Number of Sources	Area Type	Pollutant	Terrain Complexity	Required Resolution
REFINED USAGE					
Grid					
a. Region Oriented	Multiple-Source	Urban Rural	O_3, HC, CO, NO_2 (1-hour), SO_2 (3- and 24-hour), TSP	Simple Complex (Limited)	Temporal Spatial
b. Specific Source Oriented	Single-Source	Rural	O_3, HC, CO, NO_2 (1-hour), SO_2 (3- and 24-hour), TSP	Simple Complex (Limited)	Temporal
Trajectory					
a. Region Oriented	Multiple-Source	Urban	O_3, HC, CO, NO_2 (1-hour), SO_2, (3- and 24-hour), TSP	Simple	Temporal Spatial (Limited)
b. Specific Source Oriented	Single-Source	Urban Rural	O_3, HC, CO, NO_2 (1-hour), SO_2, (3- and 24-hour), TSP	Simple Complex (Limited)	Temporal Spatial (Limited)
Gaussian					
a. Long-Term Averaging	Multiple-Source Single-Source	Urban Rural	SO_2 (Annual), TSP, NO_2 (Annual)*	Simple	Spatial
b. Short-Term Averaging	Multiple-Source Single-Source	Urban Rural	SO_2 (3- and 24-hour), CO, TSP, NO_2, (1-hour)*	Simple Complex (Limited)	Temporal Spatial
REFINED/SCREENING USAGE					
Isopleth	Multiple-Source	Urban	O_3, HC, NO_2 (1-hour)	Simple Complex (Limited)	Temporal (Limited)
SCREENING USAGE					
Rollback	Multiple-Source Single-Source	Urban Rural	O_3, HC, NO_2 SO_2, CO, TSP	Simple Complex (Limited)	--
Box	Multiple-Source	Urban	O_3, HC, CO, NO_2 (1-hour), SO_2 (3- and 24-hour), TSP	Simple Complex (Limited)	Temporal

---

* Only if NO_2 is taken to be total NO_x.

Having made our general points in previous sections, however, we associate here some specific models with our generic model categories. Though this is not central to our discussion of model performance standards, it may be helpful in linking specific models to the issues and applications for which they are most suited.

In Table IV-4 we associate a number of specific models with the generic model types identified earlier. We included many of the models with which we were familiar. Because the list is intended only to be a representative one, we did not seek to make it fully complete. Many other models, particularly Gaussian ones, certainly exist and would be appropriate for use in the proper circumstances.

For the models listed in Table IV-4, a detailed summary of their characteristics is provided in Appendix B. Among the information contained there is the following: model developer, EPA recommendation status, technical description, and model capabilities. The last of these is further subdivided into source type/number, pollutant type, terrain complexity, and spatial/temporal resolution.

E. AIR QUALITY MODELS: A SUMMARY

In Chapter III we identified generic classes of air quality issues. In this chapter we defined generic types of models. Having done so, we associated the two, identifying those issues for which each model was a potentially suitable analysis tool. We also described the technical formulations employed in each generic type of model, indicating some key limitations.

As noted in Table IV-1, several generic model types may be of potential use in addressing the same generic class of issue. Only by considering the characteristics of a proposed application can a final choice of model be

TABLE IV-4. SOME AIR QUALITY MODELS

Generic Model Type	Specific Model Name
Refined Usage	
Grid	
a. Region Oriented	SAI
	LIRAQ
	PICK
b. Specific Source Oriented	EGAMA
	DEPICT
Trajectory	
a. Region Oriented	DIFKIN
	REM
	ARTSIM
b. Specific Source Oriented	RPM
	LAPS
Gaussian	
a. Long-term Averaging	AQDM
	CDM
	CDMQC
	TCM
	ERTAQ*
	CRSTER*
	VALLEY*
	TAPAS*
b. Short-term Averaging	APRAC-1A
	CRSTER*
	HANNA-GIFFORD
	HIWAY
	PTMTP
	PTDIS
	PTMAX
	RAM
	VALLEY*
	TEM
	TAPAS*
	AQSTM
	CALINE-2
	ERTAQ*
Refiner/Screening Usage	
Isopleth	EKMA
	WHITTEN
Screening Usage	
Rollback	LINEAR ROLLBACK
	MODIFIED ROLLBACK
	APPENDIX J
Box	ATDL

---

* These models can be used for both long-term and short-term averaging.

made. To facilitate the comparison between model capabilities and application requirements, we defined a set of application attributes. We then matched the two, identifying for each generic model type the combinations of application attributes for which it was suited.

In this chapter we defined the interface between issue, model, and application. In addition, we mentioned some specific air quality models within each model category, giving additional detail on each in Appendix B.

With the completion of this chapter, we are ready to consider model performance measures. In the next chapter, we identify performance measures appropriate for the consideration of each air quality issue. Having done so, we examine the interface of performance measure and model category. Finally, in Chapter VI, we discuss several alternative rationales and formats for setting model performance standards. These are designed to be consistent with the performance measures defined in Chapter V.

V MODEL PERFORMANCE MEASURES

The central purpose of this report is to identify means for setting standards for air quality model performance. As prologue to doing so, we identified generic types of air quality issues in Chapter III and generic classes of air quality models in Chapter IV, exploring their interrelationships. Now it remains to discuss the model performance measures for which performance standards must be set. Several rationales for setting these standards are presented in Chapter VI.

In this chapter our discussion proceeds as follows: We first identify generic types of performance measures; we then suggest some specific performance measures (describing them in detail in Appendix C); and finally we match generic performance measures to the issue/model/application combinations presented in earlier chapters. Before beginning, however, the notion of a model "performance measure" needs to be defined in more detail.

Typically, air quality models are used in the following context: a problem is posed, a model is chosen that is suitable for use in addressing the issue/application, existing data are assembled for input and additional data are gathered (if needed), and a simulation is conducted. Results often are expressed in the form of spatially and temporally varying concentration predictions for one or many pollutant species. Since most problems are hypothetical ones posing "what-if" questions (e.g., what if a new power plant is built, or what if population growth and development proceeds as forecast), model results in such situations are inherently nonverifiable. Consequently, before its results can be accepted, the reliability of the chosen model must be demonstrated. Most frequently, "validation" is accomplished by using the model to simulate pollutant concentrations in a test situation which is similar to the hypothetical one and for which measurement data are available. A region-oriented model (urban or regional scale) may be required to predict region-wide concentrations resulting from conditions existing on some past date. A specific-source model may have to reproduce the downwind concentrations resulting from emissions from an existing source having size and siting characteristics similar to the proposed one. If its predictions are judged to be in sufficient agreement with observed data, the model is then accepted as a satisfactory tool for use in addressing the hypothetical problem.

However, what do we mean by "satisfactory" agreement between prediction and observation? What are the quantities most appropriate for use in characterizing differences between the two? Within what range of values must these quantities remain? The values for how many different quantities must be "satisfactory" before we judge model predictions to be acceptably near test case observations?

In this chapter, we explore the second of these questions. In doing so, we identify a set of model performance measures, surrogate quantities whose values serve to characterize the comparison between prediction and observation. We match these performance measures with the generic types of air quality issues identified in Chapter III and the generic classes of air quality models listed in Chapter IV. We defer until Chapter VI the next and final step: the specification of model performance standards against which to compare for acceptability the values of the model performance measures.

A. THE COMPARISON OF PREDICTION WITH OBSERVATION

Before accepting a model for use in addressing hypothetical air quality questions, the user must validate it. This is often done by demonstrating its ability to reproduce a set of test results, usually consisting of observational concentration data recorded at a number of measurement stations for several hours during the day. In comparing predictions with observation, several questions should be asked. Among these are the following:

- What are the differences? How much does prediction differ from observation at the location of the peak concentration level and at each of the monitoring stations? What is the spatial and temporal distribution of the residuals (the difference between prediction and observation)? Do these differences correlate with diurnal changes in atmospheric characteristics (mixing height, wind speed, or solar irradiation, for instance)? If more than one species is being considered, are there differences in performance between each species?
- How serious are the differences? Are peak concentration levels widely different? Are the estimates of the area in violation of the NAAQS in substantial disagreement? How near to agreement are the estimates of the area exposed to concentrations within 10 percent of the peak value? Are differences in the timing and spatial distribution of concentrations such that the expected health impacts on the population (exposure/dosage) are of different magnitude? Do the predicted and observed patterns and levels of concentrations lead to seriously different conclusions about the required amount and cost of emissions control? Are policy decisions deriving from prediction and observation different (such as a "build-no build" decision on a power plant based on PSD considerations)?
- Are there straightforward reasons for the differences? Are the locations and timing of the concentration peaks slightly different between prediction and observation? (If concentration gradients within the pollutant cloud are steep, even a slight difference in cloud location can produce large discrepancies at set monitoring sites. Such a problem could occur if there were only slight errors in the wind speed or direction input to the model. In such an instance, model performance might otherwise

be perfectly adequate.) Are wide fluctuations in ground-level concentrations and thus station measurements produced by relatively small discrepancies between the modeled and the actual atmospheric characteristics? [This "multiplier effect" can occur downwind of an elevated point source, for example. Because the emissions plume from a point source has dimensions much greater downwind than crosswind, slight changes in the atmospheric profile (stability category), having an effect on plume rise and dispersion, have a more than proportionate effect both on the downwind distance at which the ground-level peak concentration occurs and on the amount of area exposed to a given concentration level.]

In the remainder of this chapter we discuss, first in generic terms and then in specific ones, several different types of model performance measures. While each type and variant is designed to highlight different aspects of the comparison between prediction and observation, they all address the general questions noted above. Those questions, and others like them, are the fundamental ones from which the notion of performance measures and standards derive.

B. GENERIC PERFORMANCE MEASURE CATEGORIES

In this section, we define several generic model performance measure categories, distinguishing among them on the basis of their general characteristics and the amount of information required to compute them. We also note three variants found among measures in each category. We then introduce some practical considerations which can limit the choice of performance measure. In Section C we list some of the specific measures included in the generic categories, beginning with a discussion of the fundamental differences between those designed to measure performance on a regional scale and those characterizing it on a specific-source scale. Details of these specific measures are provided in Appendix C.

1. The Generic Measures

We consider here four generic performance measure categories: peak, station, area, and exposure/dosage. The first category contains those measures related to the differences between the predicted and observed concentration peak, its level, location and timing. The second category includes measures based upon concentration differences between prediction and observation at specific measurement stations. Within the third category are contained those measures based upon concentration field differences throughout a specified area. The fourth category includes measures derived from differences in population exposure and dosage within a specified area.

Each of these generic performance measure categories requires successively greater knowledge of the spatial and temporal distribution of concentrations. We show in Figure V-1 a schematic representation illustrating several distinct levels of knowledge about regional concentrations. A similar schematic appropriate for source-specific situations is shown in Figure V-2. Listed in Table V-1 are the information requirements for the four categories. These range from an estimate of a simple scalar quantity, concentration at the peak, all the way to full knowledge of the spatially and temporally resolved concentration field and population distribution. For peak measures, the concentration residuals (the difference between predicted and observed values) are required at a single point and time. For station measures, the temporal variations of the residuals are required at several points. For both area and exposure/dosage measures, the full residual field is required, both spatially and temporally resolved. The latter type of measure requires, in addition, the spatial and temporal history of population movement within the area of interest.

As the information content increases, the ability of the performance measure to characterize the comparison between prediction and observation also can increase. However, measures from different categories tend to emphasize different aspects of the comparison. For this reason, several

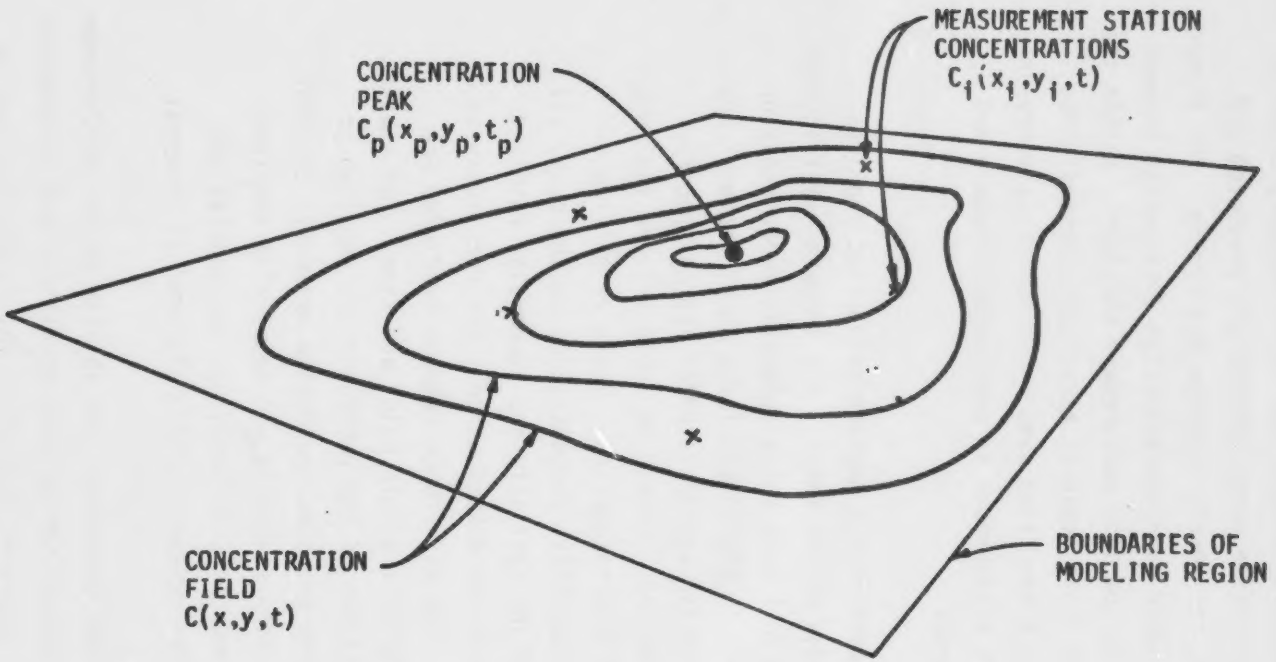

MEASUREMENT STATION CONCENTRATIONS
$C_i(x_i,y_i,t)$
CONCENTRATION PEAK
$C_p(x_p,y_p,t_p)$
CONCENTRATION FIELD
$C(x,y,t)$
BOUNDARIES OF MODELING REGION

FIGURE V-1. VARIOUS LEVELS OF KNOWLEDGE ABOUT REGIONAL CONCENTRATIONS

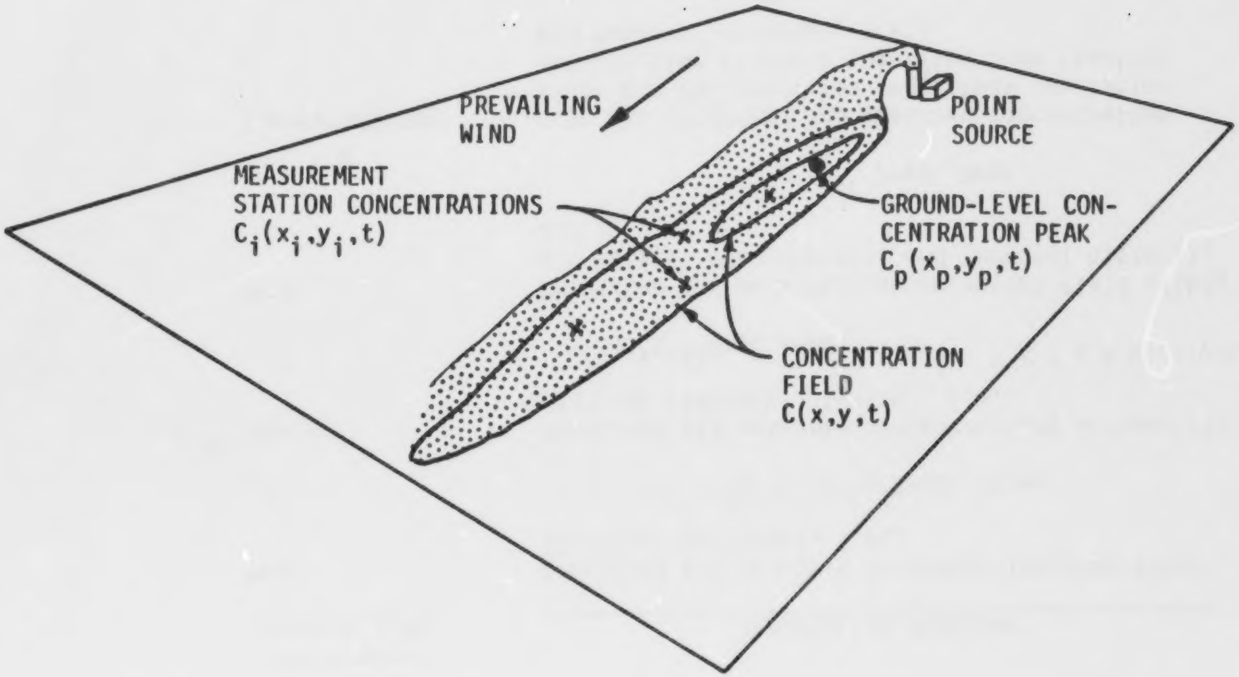

PREVAILING
WIND
POINT
SOURCE
MEASUREMENT
STATION CONCENTRATIONS
$C_i(x_i, y_i, t)$
GROUND-LEVEL CON-
CENTRATION PEAK
$C_p(x_p, y_p, t)$
CONCENTRATION
FIELD
$C(x, y, t)$

FIGURE V-2. VARIOUS LEVELS OF KNOWLEDGE ABOUT SPECIFIC-SOURCE CONCENTRATIONS

TABLE V-1. GENERIC PERFORMANCE MEASURE INFORMATION REQUIREMENTS

Generic Performance Measure Type	Information Required
Peak	Predicted and measured concentration peak (level, location, and time), i.e., $C_p(x_p, y_p, t_p)_{\text{Pred., Meas.}}$
Station	Predicted and measured concentrations at specific stations (temporal history), i.e., $C_i(x_i, y_i, t)_{\text{Pred.,Meas.}}$, $1 \leq i \leq N$ stations
Area	Predicted and measured concentration field within a specified area (spatial and temporal history), i.e., $C(x,y,t)_{\text{Pred.,Meas.}}$
Exposure/dosage	Both the predicted and measured concentration field and the predicted and actual population distribution within a specified area (spatial and temporal history), i.e., $C(x,y,t)_{\text{Pred.,Meas.}}$ $C(x,y,t)_{\text{Pred.,Actual}}$

types of performance measures are usually required in order to fully characterize a model's ability to reproduce observationally obtained data.

Because a model predicts well the observed concentration peak, for instance, does not necessarily mean its predictions can reproduce the spatially distributed concentration field. A comparison of the temporal history of concentration values at several specific stations might give a better indication of spatial model behavior. Even this might not prove conclusive. The prevailing direction of the winds input to the model might have been slightly in error. This may have little impact on concentration levels, resulting only in a pollutant cloud slightly displaced from its actual location. If concentration gradients are steep within the cloud, station predictions might not agree well with the values observed, even though the model might not be significantly deficient. In such a circumstance, area measures might provide a better means for assessing model performance. For instance, the areas in excess of a specified concentration value could be compared for several values ranging between the peak and background values.

Even employing the above measures, the degree of seriousness of the disagreement between prediction and observation might not be obvious. Since health effects result from both the pollutant level and length of exposure, measures expressing differences in exposure/dosage might give an indication of a model's ability to estimate the interaction of population with pollutant. This might be helpful in a number of circumstances. For example, suppose prevailing winds on "worst" episode days carry the pollutant cloud containing ozone and its precursors into adjacent rural areas before the early-afternoon peak occurs. If few people live in the affected area, exposure/dosage measures may indicate that the model's failure to accurately predict peak concentrations is of little practical consequence.

2. Some Types of Variations Among Performance Measures

Three types of variations are found among performance measures: scalar, statistical, and "pattern recognition." Those measures of the first type are based upon a comparison of the predicted and observed values of a specific quantity: the peak concentration level, for instance. Those of the second type compare the statistical behavior (the mean, variance and correlation, for example) of the differences between the predicted and observed values for the quantities of interest. Measures of the final type are useful in providing qualitative insight into model behavior, transforming concentration "residuals" (the differences between predicted and observed values) into forms that highlight certain aspects of model performance and thus triggering "pattern recognition."

In order to illustrate the types of variations found in each generic performance measure category, we present Table V-2. Some typical examples are included for each category/variation combination. In section D of this chapter, a number of specific performance measures are listed. Examined in detail in Appendix C, they are classified according to the scheme presented here.

3. Several Practical Considerations

Several practical considerations have a strong impact on the choice of model performance measures. Each of these derive from limitations on the degree of spatial resolution attainable with most models and measurement networks.

Ideally, in assessing the performance of a model, one might want to examine for several hours during the day the agreement between prediction and observation throughout the concentration field (the spatial distribution of concentrations). Differences between the predicted and observed values of the following could be uncovered thereby: the location, timing, and level of the concentration peak; the area exposed to a concentration in excess of a given value (e.g., the NAAQS); and the concentration values at stations within a measurement network.

TABLE V-2. TYPES OF VARIATIONS AMONG GENERIC PERFORMANCE MEASURE CATEGORIES

Generic Performance Measure Category	Types of Variations	Typical Example
Peak	Scalar	Concentration residual* at the peak.
	Pattern recognition	Map showing locations and values of maximum one-hour-average concentrations for each hour.
Station	Scalar	Concentration residual at the station measuring the highest value.
	Statistical	Expected value, variance and correlation coefficient of the residuals for the modeling day at a particular measurement station.
	Pattern recognition	At the time of the peak (event-related), the ratio of the residual at the station having the highest value to the average of the residuals at the other station sites (this can indicate whether the model performs better near the peak than it does throughout the rest of the modeled region).
Area	Scalar	Difference in the fraction of the modeled area in which the NAAQS are exceeded.
	Statistical	At the time of the peak, differences in the area/concentration frequency distribution.
	Pattern recognition	For each modeled hour, isopleth plots of the ground-level residual field.

---

*Residual: The difference between "predicted" and "observed."

TABLE V-2 (Concluded)

Generic Performance Measure Category	Types of Variations	Typical Example
Exposure/dosage	Scalar	Differences in the number of person-hours of exposure to concentrations greater than the NAAQS.
	Statistical	Differences in the exposure concentration frequency distribution.
	Pattern recognition	For the entire modeled day, an isopleth plot of the ground level dosage residuals.

Difficulties hindering such an examination arise from two sources: the limited spatial resolution of the model and the sparsity of the measurement network. While some models, such as the Gaussian ones, are analytic and thus able to resolve the concentration field, many cannot do so completely. Grid models, for example, predict a single average concentration value for each cell. For this reason, they can not resolve the concentration field on a spatial scale any finer than the intergrid spacing (usually on the order of one or two kilometers for urban scale grid models). Trajectory models are similarly limited: They can resolve the concentration field only as finely as the dimensions of the air parcel being simulated. Further, predictions are computed only for a particular space-time track, and not for the entire concentration field.

The relatively small number of stations in most measurement networks limits the ability to reconstruct completely the concentration field actually occurring on the modeled day. While stations are well-placed in some networks, in others they are not. Thus, not only are stations often 3-10 kilometers apart, their placement does not always guarantee the observation of peak or near-peak concentrations. Further, even in extended urban areas, seldom does the number of stations exceed 10 to 20.

For these reasons, concentration fields generally are not known with precision, from either model predictions or observational data. Estimates of the spatial distribution of concentrations can be obtained only by inference from "sparse" data. The use of numerical processes, such as interpolation and extrapolation, to extend that data introduces additional uncertainty into the comparison of predictions with observations.

Another consequence results from the limited resolution of measurement networks: The value of the concentration peak actually occurring on the day of observation may not be known. Measurement networks usually consist of fixed stations arranged in a set pattern. Unless the air parcel containing the peak drifts over or near one of the stations, the maximum concentration value sensed by the network will be less (sometimes substantially so) than the value of the actual maximum. When prevailing winds and pollutant chemistry are highly predictable for the days of worst episode conditions, station placement can be designed so as to maximize the likelihood of sensing the true peak. When conditions are not so predictable, a measurement network with a modest number of stations has little chance of "seeing" the true peak. For instance, suppose the cloud containing the peak and all concentrations within 20 percent of it covers an area of 25 square kilometers in an urban area having a total area of 1000 square kilometers. If the cloud has an equal likelihood of being above any point in the urban region at the time of the peak, by dividing the area of the cloud into the total urban area, we can make a crude estimate of the number of stations required to guarantee a measurement within 20 percent of the peak: 40 stations evenly spaced about 5 kilometers apart throughout the urban region would be required. Even if the probable location of the cloud were known to be within an area equal to one-quarter of the urban area, 10 stations would be required just within that small area. This degree of station density is high and may not be found in many circumstances.

The above example is a simplistic one. The design of actual station placement can be a far more complex process than indicated here. However, the example serves to underscore the main point: a measurement network, though satisfying EPA regulations,* may still be unable to guarantee an observation "close" to the actual concentration peak, i.e., within 10 to 20 percent.

The points raised in the above discussion have some practical implications for the choice of a model performance measure. Among these are the following:

---

* Source: 40 CFR §51.17 (1975).

> Performance measures relying on a comparison of the predicted and "true" peak concentrations may not be reliable in all circumstances since measurement networks can provide only the concentration at the station recording the highest value, not necessarily the value at the "true" peak.

> Performance measures relying on a comparison of the predicted and "true" concentration fields may not be computationally feasible since neither predicted nor "true" concentration fields are always resolvable, spatially or temporally, at the scales required for comparison.

> Performance measures based upon a comparison of predicted and "true" exposure/dosage, though they are appealing because of their ability to serve as surrogates for the health effects experienced by the populace, may not be computationally feasible because of the difficulty in measuring the "true" population distribution and the "true" concentration field. (We do suggest in Chapter VI, however, one means by which health effects considerations can be accounted for implicitly.)

> Performance measures based upon a comparison of the predicted and observed concentrations at station sites in the measurement network may be of the greatest practical value.*

While the above points are general ones, exceptions to them do occur in specific applications. Also, certain performance measures, though not fully reliable on their own, can be useful in a qualitative sense when used in conjunction with other measures.

C. A BASIC DISTINCTION: REGIONAL VERSUS SOURCE-SPECIFIC PERFORMANCE MEASURES

Some models are used to address multiple-source, region-oriented issues; others are applied to consider single-source issues. The performance measures appropriate for each differ. We consider here the distinction between regional and source-specific performance measures.

---

*Note caveat on pages VI-18 and VI-19, with respect to point source applications.

The distinction is drawn not so much between the type of performance measure used (peak, station, area, or exposure/dosage), but rather between the spatial scales over which it is applied. To address urban or regional scale issues (SIP/C, AQMP), we must consider a region hundreds of square kilometers in area, with the spatial and temporal distribution of concentrations the result of emissions from many sources. The quantities of interest are: the regional peak concentration (its level, location and timing) and for each hour during the day (particularly at the time of the peak), the spatial distribution of the pollutant concentrations, by species. This information is frequently conveyed in the form of a concentration isopleth diagram, an example of which is shown in Figure V-3. The diagram shown was produced by the SAI Urban Airshed Model, illustrating its ozone predictions for the Denver Metropolitan region at Hour 1200-1300 MST on 29 July 1975.

To address single-source issues, on the other hand, we consider only the region downwind of the specific source being modeled. While emissions from it contribute to the overall pattern and level of regional pollutant concentrations, it is usually the incremental impact of those emissions that are of concern. The principal quantities of interest are: the peak incremental ground-level concentration downwind of the source and the spatial distribution of the incremental concentrations within the downwind ground-level "footprint." Specific parameters describing the latter are: the area within which concentrations exceed a certain value and the shape of the concentration isopleths, usually conveyed in the form of a diagram such as the one shown in Figure V-4. This diagram was constructed using a Gaussian formulation for a continuously emitting elevated point source. Conditions are in steady-state and "perfect" reflection from the ground is assumed. No inversion layer exists. It should be noted that winds are unlikely to persist long enough for actual conditions ever to resemble these isopleths beyond 20 to 25 km (about 6 to 10 hours).

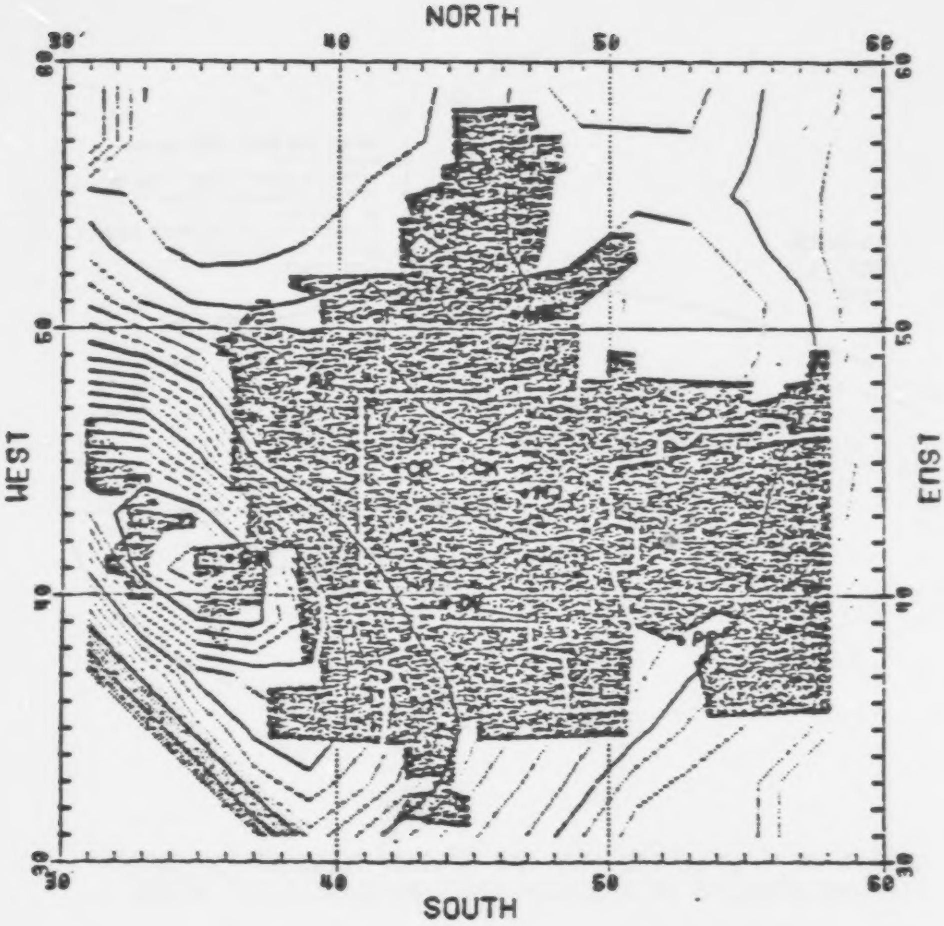
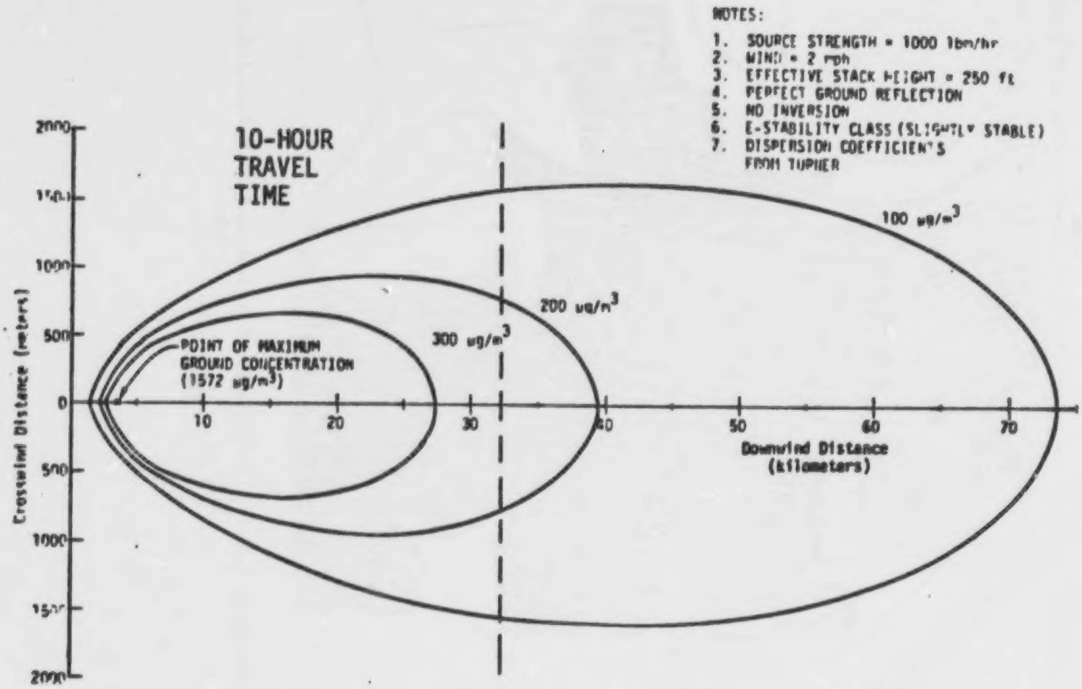

NORTH
30
40
50
60
WEST
EAST
60
50
40
30
SOUTH

FIGURE V-3. SAMPLE REGIONAL ISOPLETH DIAGRAM ILLUSTRATING OZONE CONCENTRATIONS (pphm) IN DENVER ON 29 JULY 1975 FOR HOUR 1200-1300 MST. Isopleth interval is 1 pphm. Grid divisions are in miles. Actual air quality monitoring station locations are shown. (This is from Anderson et al., 1977.)



NOTES:

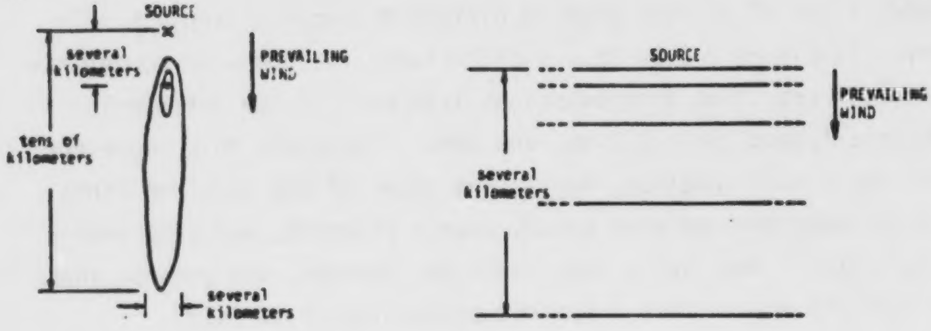
1. SOURCE STRENGTH = 1000 lbm/hr
2. WIND = 2 mph
3. EFFECTIVE STACK HEIGHT = 250 ft
4. PERFECT GROUND REFLECTION
5. NO INVERSION
6. E-STABILITY CLASS (SLIGHTLY STABLE)
7. DISPERSION COEFFICIENTS
   FROM TURNER

FIGURE V-4. SAMPLE SPECIFIC-SOURCE ISOPLETH DIAGRAM ILLUSTRATING CONCENTRATIONS DOWNWIND OF A STEADY-STATE GAUSSIAN POINT SOURCE

Other types of sources produce different downwind isopleth patterns. In Figure V-5 we show qualitatively the downwind concentration patterns resulting from emissions from each of the three principal source types: point, line, and area. These are only representations; the actual location, level, and shape of the isopleth lines are heavily dependent on wind speed, source strength, and atmospheric stability class. The figure does indicate, however, the general shape of the downwind area within which the source impact is felt.

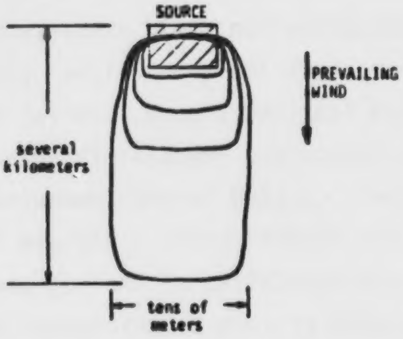
The type of source provides information in two areas: It identifies the modeling region within which the peak, station, area, and exposure/dosage performance measures are to be applied; and it provides insight for monitoring network design. The observational data against which model performance is to be judged are gathered at the measurement stations within that network. To measure properly the impact produced by a specific source, the measurement network should be deployed in a pattern consistent with the concentration field shapes shown in Figure V-5. The station designed to measure the ground-level peak concentration should be located downwind from the source, several kilometers distant for an elevated point source and immediately adjacent for either a line or an area source. Located farther downwind are those stations designed to resolve the concentration field and to determine the concentration value most representative of the regional incremental impact of the source. A schematic of such a measurement network for a point source is presented in Figure V-6, showing one possible configuration for the stations.

Several difficulties arise in practice: Wind direction is changeable, and the location of ground-level footprints is very sensitive to atmospheric stability. These problems are particularly acute when the emitter being considered is an elevated point source. To illustrate, we show in Figure V-7 the locus of the downwind footprint if all wind directions are considered equally likely to occur. If we idealize the concentration isopleths as being elliptical in shape, we can determine an


SOURCE
several kilometers
tens of kilometers
several kilometers
PREVAILING WIND
SOURCE
PREVAILING WIND
several kilometers

(a) Point Source (e.g., Power Plant)

(b) Line Source (e.g., Traffic Link)


SOURCE
PREVAILING WIND
several kilometers
tens of meters

(c) Area Source (e.g., Parking Lot)

FIGURE V-5. CONCENTRATION ISOPLETH PATTERNS FOR VARIOUS SOURCE TYPES. Continuous and uniform emissions are assumed. Steady-state conditions prevail.

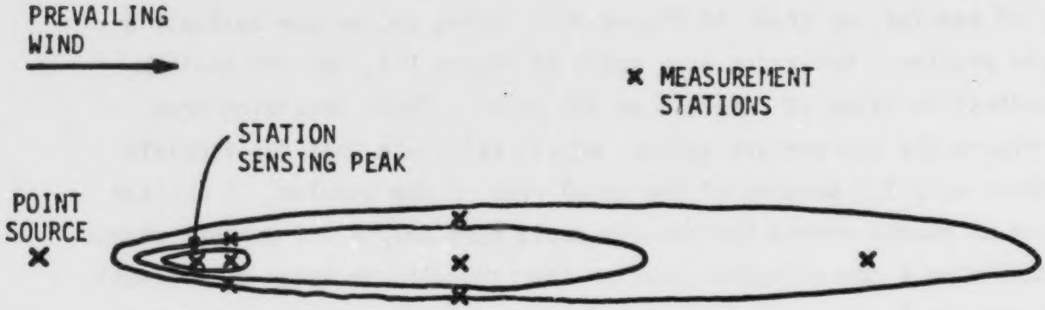

PREVAILING WIND
MEASUREMENT STATIONS
STATION SENSING PEAK
POINT SOURCE

FIGURE V-6. SCHEMATIC OF A POINT SOURCE MEASUREMENT NETWORK

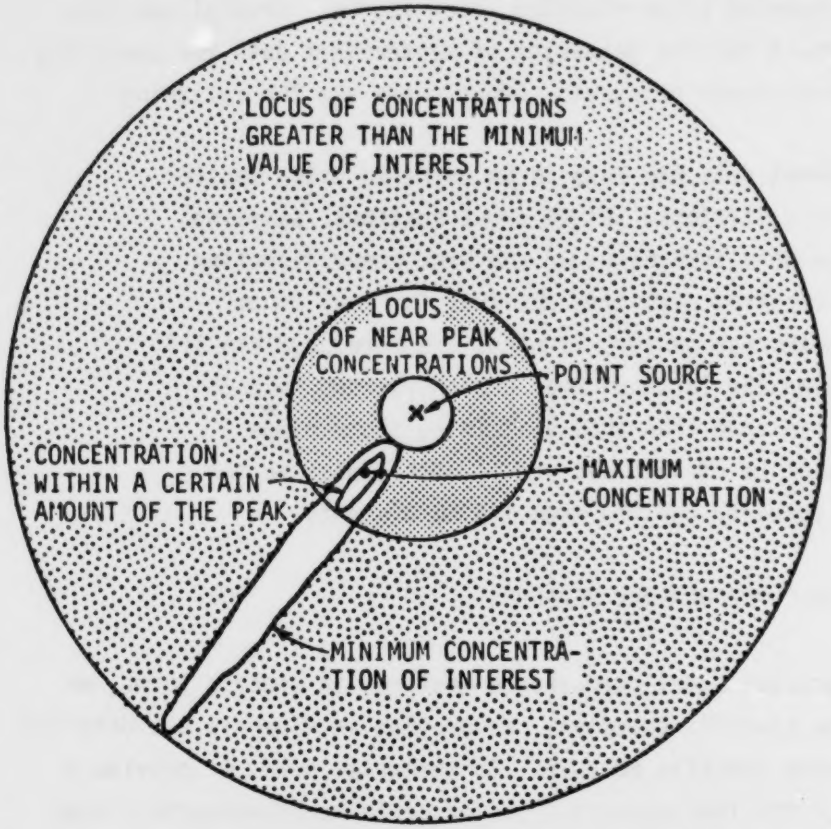

LOCUS OF CONCENTRATIONS GREATER THAN THE MINIMUM VALUE OF INTEREST
LOCUS OF NEAR PEAK CONCENTRATIONS
POINT SOURCE
CONCENTRATION WITHIN A CERTAIN AMOUNT OF THE PEAK
MAXIMUM CONCENTRATION
MINIMUM CONCENTRATION OF INTEREST

FIGURE V-7. LOCUS OF POSSIBLE FOOTPRINT LOCATIONS FOR AN ELEVATED POINT SOURCE. All wind directions are considered equally likely.

expression for the ratio of the area within a given isopleth to the area of annulus, as shown in Figure V-7. Doing so, we can evaluate a sample problem. Referring once again to Figure V-4, let the minimum concentration value of interest be 300 $\mu g/m^3$. Then, obtaining from the figure the appropriate values, we can calculate that the isopleth contains only 1.2 percent of the total area of the annulus. A monitor placed at random within the annulus would have only a 1.2 percent chance of observing a concentration greater than the minimum value of interest. This problem is compounded if we consider variations in the inner and outer radii due to the varying dispersive power of the wind.

The message of all this is clear: When winds are variable, fixed monitoring stations have little chance of characterizing the concentration field downwind of an elevated point source. Several specific implications result for the gathering of measurement data for computing point source performance measures. Among these are the following:

> Measurement data may have to be gathered using mobile monitoring stations. Plume cross sectional sampling could be done then based on the wind speed/direction and atmospheric stability observed in "real time."

> The annulus (or sector, if winds are more predictable) containing the locus of peak concentrations is much smaller in area than that containing the minimum concentration of interest and is much closer to the source (usually ranging from 1-5 km distant).

D. SOME SPECIFIC PERFORMANCE MEASURES

Having discussed model performance measures in generic terms, we now present some specific examples. We provide in Appendix C a detailed discussion of each specific measure. To summarize here, we provide a list for each of the four generic types of performance measures: peak (Table V-3), station (Table V-4), area (Table V-5), and exposure/dosage

TABLE V-3. SOME PEAK PERFORMANCE MEASURES

Type	Performance Measure
Scalar	a. Difference[*] in the peak ground-level concentration values.
	b. Difference in the spatial location of the peak.
	c. Difference in the time at which the peak occurs.
	d. Difference in the peak concentration levels at the time of the observed peak.
	e. Difference in the spatial location of the peak at the time of the observed peak.
Pattern recognition	Map showing the locations and values of the predicted maximum one-hour-average concentrations for each hour.

---

[*] "Difference" as used here usually refers to "prediction minus observation."

TABLE V-4. SOME STATION PERFORMANCE MEASURES

Type	Performance Measure
Scalar	a. Concentration residual at the station measuring the highest concentration (event-specific time and fixed-time comparisons).
	b. Difference in the spatial locations of the predicted peak and the observed maximum (event-specific time and fixed-time comparisons).
	c. Difference in the times of the predicted peak and the observed maximum.
Statistical	a. For each monitoring station separately, the following concentration residuals statistics are of interest for the entire day: 1) Average deviation 2) Average absolute deviation 3) Average relative absolute deviation 4) Standard deviation 5) Correlation coefficient 6) Offset-correlation coefficient.
	b. For all monitoring stations considered together, the following residuals statistics are of interest: 1) Average deviation 2) Average absolute deviation 3) Average relative absolute deviation 4) Standard deviation 5) Correlation coefficient 6) Estimate of bias as a function of concentration 7) Comparison of the probabilities of concentration exceedances as a function of concentration
	c. Scatter plots of all predicted and observed concentrations with a line of best fit determined in a least squares sense.
	d. Plot of the deviations of the predicted versus observed points from the perfect correlation line compared with estimates of instrumentation errors.

TABLE V-4 (Concluded)

Type	Performance Measure
Pattern recognition	a. Time history for the modeling day of the predicted and observed concentrations at each site. b. Time history of the variations over all stations of the predicted and observed average concentrations. c. At the time of the peak (event-related), the ratio of the normalized residual at the station having the highest value to the average of the normalized residuals at the other stations.

TABLE V-5. SOME AREA PERFORMANCE MEASURES

Type	Performance Measure
Scalar	a. Difference in the fraction of the area in which the NAAQS are exceeded.
	b. Nearest distance at which the observed concentration is predicted.
	c. Difference in the fraction of the area in which concentrations are within 10 percent of the peak value.
Statistical	a. At the time of the peak, differences in the fraction of the area experiencing greater than a certain concentration; differences in the following are of interest: 1) Cumulative distribution function 2) Density function 3) Expected value of concentration 4) Standard deviation of density function
	b. For the entire residual field, the following statistics are of interest: 1) Average deviation 2) Average absolute deviation 3) Average relative absolute deviation 4) Standard deviation 5) Correlation coefficient 6) Estimate of bias as a function of concentration 7) Comparison of the probabilities of concentration exceedances as a function of concentration
	c. Scatter plots of prediction-observation concentration pairs with a line of best fit determined in a least squares sense.
Pattern recognition	a. Isopleth plots showing lines of constant pollutant concentration for each hour during the modeling day.
	b. Time history of the size of the area in which concentrations exceed a certain value.
	c. Isopleth plots showing lines of constant residual values for each hour during the day ("subtract" prediction and observed isopleths).
	d. Isopleth plots showing lines of constant residuals normalized to selected forcing variables (inversion height, for instance).
	e. Peak-to-overall performance indicator, computed by taking the ratio of the mean residual in the area of the peak (e.g., where concentrations are within 10 percent of the peak) to the mean residual in the overall region.

(Table V-6). We include scalar, statistical, and qualitative/composite pattern recognition variants.

E. MATCHING PERFORMANCE MEASURES TO ISSUES AND MODELS

To this point we have identified several performance measures categories, discussed their general attributes and data requirements, and associated with them a number of specific performance measures. Two tasks remain in this chapter: We first indicate for each of the generic types of issues the performance measures most appropriate for use; we then discuss the capability of each generic class of model to calculate those measures.

1. Performance Measures and Air Quality Issues

In Chapter III we identified seven generic types of air quality issues, dividing them into two broad categories. Within the first of these multiple-source issues, we included: State Implementation Plan/ Compliance (SIP/C) and Air Quality Maintenance Planning (AQMP). The second category, source-specific issues, was defined to contain the following: Prevention of Significant Deterioration (PSD), New Source Review (NSR), Offset Rules (OSR), Environmental Impact Statements/ Reports (EIS/R), and Litigation (LIT). For each of these issues we now consider some important distinctions that bear on the selection of the most appropriate model performance measures (PMs).

> Multiple-Source Issues

- SIP/C. The compliance portion of a SIP details plans for achieving ambient pollutant levels at or below the NAAQS in Air Quality Control Regions (AQCRs) currently in noncompliance. Because it is the peak concentration level that is of primary concern, a model should demonstrate its ability to predict that peak. For a day chosen as the one

TABLE V-6. SOME EXPOSURE/DOSAGE PERFORMANCE MEASURES

Type	Performance Measure
Scalar	a. Difference for the modeling day in the number of person-hours of exposure to concentrations: 1) Greater than the NAAQS 2) Within 10 percent of the peak. b. Difference for the modeling day in the total pollutant dosage.
Statistical	a. Differences in the exposure/concentration frequency distribution function; differences in the following are of interest: 1) Cumulative distribution function 2) Density function 3) Expected value of concentration 4) Standard deviation of density function b. Cumulative dosage distribution function as a function of time during the modeled day.
Pattern recognition	For each hour during the modeled day, an isopleth plot of the following (both for predictions and observations): 1) Dosage 2) Exposure

to be used for model verification, peak performance measures should be computed. Also contained within SIPs are emissions control strategies. To assess the effects of controlling specific sources, a model must be capable of spatially resolving its concentration predictions. Area PMs should be calculated, if possible, to evaluate a model's ability to do so. Station PMs are another means to evaluate model spatial resolution, although pollutant cloud offset can account sometimes for apparent large discrepancies. Because SIP/C is most frequently an issue in densely populated urban areas, large differences in health effect impact can exist between prediction and observation. Exposure/dosage PMs should be calculated, if possible, in order to evaluate the acceptability of a model's performance.

- AQMP. Detailed within the maintenance portion of a SIP are procedures for insuring, once compliance has been achieved, that ambient pollutant concentrations do not again rise above the NAAQS. Because violation of the NAAQS is an issue, peak PM's are important measures of model performance. However, because pollutant levels are low (relative to the values before compliance), small errors in model performance might not produce a large uncertainty in expected health impact. Consequently, the use of exposure/dosage PMs may not be necessary. Also, emissions control strategies may not be as global. Retrofit of control devices on existing sources will have been accomplished. Automotive emissions will have been controlled (presumably) such that point sources will contribute a large fraction of the emissions inventory. While incremental growth and development will alter the spatial and temporal distribution of pollutants, the need for modeling spatial resolution may not be so crucial as it was with SIP/C. Agreement between prediction and observation as measured by area and station PMs, while desirable, may not always be required within the same tolerance as for SIP/C issues.

> Specific-Source Issues

- PSD. Individual sources are not permitted to cause more than small incremental increases in concentrations in areas currently in attainment of the NAAQS. Since these so-called "Class I" regions (often state or national parks) are generally some distance from the polluting source (>10 kilometers), a model must be able to predict accurately ground-level concentrations some distance downwind from the source. If the source being modeled is by itself likely to produce near-stack ground-level concentrations in excess of the NAAQS or increments greater than Class II allowable increments, peak measures are of particular interest. Otherwise, "far-field" concentration predictions are more important than estimates of the peak value. Downwind station PMs are often the measures most suitable for evaluating model predictions for PSD Class I. Also, plumes from point source are very narrow, that is, their cross-wind dimensions are much smaller than their downwind ones. Consequently, the incidence of a Class I violation may be quite sensitive to model performance, as measured by area PMs. However, exposure/dosage PMs are not likely to be of interest because of the sparsity of population in areas where PSD is an issue and the relatively low concentrations occurring there.
- NSR. New source review is an important issue in both urban and nonurban regions. With the density of population in urban areas, many persons may live within a short distance (<5 kilometers) of a source. The ground-level peak concentration, then, may be an important

indicator of near-source health impact. Prediction of that peak, as measured by a peak PM, may be an important model performance requirement. However, because ground-level concentrations fall off rapidly farther downwind and because of the "narrowness" of the plume, differences in exposure and dosage between prediction and observation may not be of substantial consequence. Close agreement, as measured by area and exposure/dosage PMs, may not be required. Also, in order to assess the impact of a new or modified source, it is necessary to know its incremental effect on regional air quality. This is best represented by an "average" concentration value (including background) well downwind of the source (>10 kilometers). Thus, a model should demonstrate its ability to reproduce measurement data at that downwind range. The use of station PMs is indicated.

- OSR. In order to construct a new source or modify an existing one in a region experiencing concentrations in excess of the NAAQS, the owner of the source must arrange for the removal of existing sources. An amount greater than the emissions from the proposed new source must be removed from the regional inventory. Currently, these "offsets" are made on the basis of emissions rather than as a result of their impact on ambient concentrations. In such a case, no air quality predictions are required (unless a region-wide violation is attributable to the source being removed or cleaned up). Only an accurate emissions inventory is necessary. However, if offsets were "negotiated" at the level of ambient concentrations, the predictions of air quality models would assume significance. The "far" downwind concentration value, representative of its regional incremental impact, would be the quantity of greatest interest, since it would describe the source's offset "potential." Station PMs then would be of use in evaluating model performance.

- EIS/R. Projects having a significant, adverse impact on air quality usually are presented for public review by means of an EIS or an EIR. Such projects generally consist of one or a few distinct sources, although some consist of a greater number. An example of the latter is the Denver Metropolitan Wastewater Overview EIS recently completed by Region VIII of the EPA. Federal funding for twenty-two separate sewerage treatment facilities was conditioned upon favorable review of the EIS which examined their combined regional impact. If the sources are widely distributed throughout the modeling region, spatial resolution may be an important model requirement. In such a case, area and station PMs would provide a useful means to verify model acceptability. If the combined emissions from the proposed sources are relatively low or they are localized to a narrow downwind plume, their incremental health impact may be small. Exposure/dosage PMs might be applied to assess model performance. However, if, as in Denver, the potential impact is more serious and widespread, this latter type of PM can be useful.
- LIT. Court challenges can arise to the basic air pollution laws themselves, to their implementation to federal regulation, or to decisions regarding specific sources (requests for variances and applications for construction/modification approval, for example). While challenges of the first two types can and have had important consequences, we identify the third type as the principal variant included in LIT. When the specific source in question

is to be located in an urban area, the model used to estimate its effects should be expected to predict both its near-source, ground-level concentration peak and its far-field "average" value. Peak and station PMs should be used. If the source is to be constructed in a rural area, PSD may be an issue in arriving at a build/no-build decision. If so, accuracy of spatial resolution could be important. The use of area PMs could be of assistance.

We summarize in Table V-7 many of the points mentioned above. In it issues are associated with the generic categories of performance measures most commonly required for use in assessing model performance. However, exceptions do occur. For this reason, the final choice of performance measures should be dictated by the character of the specific application.

2. Performance Measures and Air Quality Models

In the previous section we associated performance measures with generic types of issues. We now discuss the ability of generic classes of models to generate predictions in a form suitable for calculation of those measures. All model types produce estimates of the concentration peak. Some can predict station concentrations. Fewer can spatially resolve the concentration field. Fewer still are able to determine an estimate of exposure/dosage. For each generic model category, we outline here their general capabilities.

> Grid. The formulation of grid models permits the estimation of concentrations averaged for each grid cell. Consequently, the concentration field can be resolved spatially as finely as the dimensions of the grid cell. The peak is estimated to be the maximum ground-level grid cell concentration occurring during the modeling day. The location of the peak is predicted only as closely as

TABLE V-7. PERFORMANCE MEASURES ASSOCIATED WITH SPECIFIC ISSUES

Issue	Performance Measure Type			
	Peak	Station	Area	Exposure/Dosage
Multiple-source				
SIP/C	X	X	X	X
AQMP	X	X	X	
Specific-source				
PSD	X	X	X	
NSR	X	X	X	
OSR		X	X	
EIS/R	X	X	X	X
LIT	X	X	X	

a single grid cell dimension. The value at the peak is predicted only as an area average in the vicinity of the peak (within one grid cell). Because of its spatial and temporal resolution, predictions suitable for calculation of station, area and exposure/dosage performance measures also can be generated.

> Trajectory. Because a single air "column" is simulated, only concentrations along the space-time track followed by the advecting air parcel can be estimated. Such models, as a consequence, can predict station concentrations only for those over which they pass. If several adjoining parcels are modeled, predictions at other stations can be determined. The spatial location of the peak can be estimated only as closely as the dimensions of the air column. The peak level is estimated to be the greatest column-averaged concentration occurring during the modeling day. Averaging can take place over the entire vertical region from the ground to the inversion base or over the lowest of several vertical column-layers. Because of their limited spatial resolution, regional trajectory models do not generate predictions in a form suitable for the calculation of area or exposure/dosage PMs. Specific-source trajectory models, on the other hand, may do so. Concentrations are predicted as a function of downwind distance from the source. Though lateral resolution is limited, concentration estimates can be put in a form appropriate for calculation of station, area and exposure/dosage PMs.

> Gaussian. Concentration field predictions are expressed analytically. Thus, subject to the steady-state limitations of their formulation, the short-term averaging versions of these models can provide their estimates in a form that is suitable for the calculation of all performance measure types. The long-term averaging versions, however,

predict regional or sector-averaged estimates of annual concentrations. Estimates of exposure/dosage (except crudely on the basis of an annual concentration level) are difficult to derive. Predictions of annual station averages, though, can be obtained for regional models of this type.

> Isopleth. Estimates in no other form than the regional peak concentration can be obtained with this method. This can be done only when the isopleth diagrams can be interpreted in an absolute sense. This is the case only when the isopleth diagram has been derived for ambient conditions similar to the ones in the area being modeled. In addition, a prediction of the peak can be verified only if a historical data base exists that is sufficient to determine a peak concentration in a previous base year and a record of the emissions cutbacks occurring since then.

> Rollback. The only prediction obtainable from rollback is an estimate of the regional peak concentration. This is determinable only if an historical data base exists such as that described for the isopleth method.

> Box. A prediction of the regional peak concentration can be determined using this method. No other estimates requiring finer spatial resolution can be computed. Diurnal variation in the estimates of regional average concentration, however, can be made.

We summarize in Table V-8 many of the points mentioned above. In it, we indicate for each generic model the type of performance measure that may be calculated, given the capabilities and limitations of each formulation.

F. PERFORMANCE MEASURES: A SUMMARY

In this chapter we identified generic performance measure categories, listed some specific performance measures, and then associated the

TABLE V-8. PERFORMANCE MEASURES THAT CAN BE CALCULATED BY EACH MODEL TYPE

	Performance Measure Type			
Model	Peak	Station	Area	Exposure/ Dosage
Refined usage				
Grid				
Region oriented	X	X	X	X
Specific source oriented	X	X	X	X
Trajectory				
Region oriented	X	X		
Specific source oriented	X	X	X	X
Gaussian				
Long-term averaging	X	X	X	
Short-term averaging	X	X	X	X
Refined/screening usage				
Isopleth	X			
Screening usage				
Rollback	X			
Box	X			

generic measure with generic issues, noting for each model type the PMs they are capable of calculating. Having done so, we are now ready to proceed with the final objective of this report: the discussion of model performance standards. The presentation in Chapter VI will be based upon the points raised in this chapter. The following are of crucial importance:

> Measurement networks often do not sense the "true" concentration peak.

> Only performance measures based upon station measurement data may be computationally feasible.

> Model predictions are often resolvable on a finer scale than measured concentrations; even though strict comparison of prediction with observation through some computed measure may not be fruitful, the model predictions themselves may still offer valuable insight.

VI MODEL PERFORMANCE STANDARDS

The central purpose of this report is to suggest means for setting performance standards for air quality dispersion models. Toward that end our discussion has proceeded as follows: Issues were identified (Chapter III); issue/model combinations were presented (Chapter IV); and alternative issue/model/performance measure associations were discussed (Chapter V). We are now at the final step: the setting of standards. To place this in the proper framework, we first identify five attributes of desirable model performance, showing how their relative importance depends on the issue being addressed and the pollutant being considered. Then we recommend specific performance measures whose values reveal the presence or absence of each performance attribute. We detail several rationales for establishing standards for those measures. To illustrate the use of these measures in assessing model performance, we present a sample case. It is based upon SAI experience in using a grid-based photochemical model in the Denver metropolitan region. Finally, we detail possible forms the actual standard might assume, suggesting a sample draft outline and format.

The subject addressed in this report is a broad and complex one. Seldom can a rule for judging model performance be stated that does not have several plausible exceptions to it. Consequently, we view the establishment of model performance standards to be a pragmatic and evolutionary exercise. As we gain experience in evaluating model performance, we will need to modify both our choice of performance measures and the range of acceptable values we insist on. Nevertheless, the process must begin somewhere. The recommendations contained in this chapter represent such a beginning.

We feel the measures and standards we suggest for use here will almost certainly change as experience improves our "collective judgment" about what constitutes model acceptability and what does not. Perhaps the number of measures will increase to provide richer insight into model performance, or perhaps the number will shrink without any loss of "information content." Regardless of the list of measures and their standards that ultimately emerges for use, it is the conceptual structuring of the performance evaluation itself that seems to be most important at this point. We must identify the attributes of a well-performing model, and we need to understand how we assess their relative importance, depending on the issue we are addressing and the pollutant species we are considering. The discussion in this chapter offers a conceptual structure for "folding in" all these concerns and suggests candidate measures and standards.

A. PERFORMANCE STANDARDS: A CONCEPTUAL OVERVIEW

The chief value of air quality models lies in their predictive ability. Only through their use can the consequences of pollution abatement alternatives be assessed and compared. Only by means of model predictions can the impact of emissions from newly proposed sources be estimated and evaluated for acceptability. However, because the questions typically asked of models are hypothetical ones, their predictions are inherently nonverifiable. Only after the proposed action has been taken and the required implementation time elapsed will measurement data confirm or refute the model's predictive ability.

Herein lies the dilemma faced by users of air quality models: If a model's predictions at some future time cannot be verified in advance, on what basis can we rely on that model to decide among policy alternatives? In resolving this, most users have adopted a pragmatic approach: If a model can demonstrate its ability to reproduce for a similar type of application a set of "known" results, then it is judged an acceptable predictive tool. It is on this basis that model "verification" has become an essential prelude to most modeling exercises.

A further difficulty exists. What constitutes a set of "known" results? This is not a problem easily solved. For "answers" to be known exactly, the "test" problem must be simple enough to be solved analytically. Few problems involving atmospheric dynamics are so simple. Most are complex and nonlinear. For these, the analytic test problem is an unacceptable one. Another, more practical alternative often is employed. For regional, multiple-source applications, the "known" results are taken to be the station measurements of concentrations actually recorded on a "test" date. For pollutants having a short-term standard, the duration of measurement is a day or less. For those subject to a long-term (annual) standard, the duration is a year or more.

For source-specific applications, the source of interest may not yet exist, permission for its construction being the principal issue at hand. For these applications, it is often necessary to verify a model using the most appropriate of several protypical "test cases." These could be assembled from measurements taken at existing sources, the variety of source size, type and location spanning the range of values found in applications of interest.

The term "known" is used imprecisely when referring to a set of measurement data. Station observations are subject to instrumentation error. The locations of fixed monitoring sites may not be sufficiently well distributed spatially to record data fully characterizing the concentration field and its peak value. Nevertheless, despite those shortcomings, "observed" data often are regarded as "true" data for the purposes of model verification.

Having assembled two sets of data, one "known" and the other "predicted," we can assess model performance by comparing one with the other. Prediction and observation, however, can be compared in many ways. We must select the quantities that can best characterize the distribution of pollutants in the ambient air, for it is through comparison of their predicted and observed ("known") values that we specify model performance. We catalogued a number of useful performance measures in Chapter IV, as well as in Appendix C. Later in this Chapter we indicate that subset we view as having the greatest practical usefulness.

Once we have decided on the performance measures best suited to our issue/application (and most feasible computationally), we can calculate these values. Having done so, however, we must ask a central question: How close must prediction be to observation in order for us to judge model performance as acceptable? In order for us to answer "how good is good," performance standards for these measures must be set, with allowable tolerances (predicted values minus observed ones) derived based upon a reasonable rationale (health effects or pollution control cost considerations, for instance).

By setting these standards explicitly, certain benefits may be gained. Among these are the following:

- > A degree of uniformity is introduced in assessing model reliability.
- > The impact of limitations in both data gathering procedures and measurement network design can be made more explicit, facilitating any review of them that may be required.
- > The performance expected of a model is stated clearly, in advance of the expenditure of substantial analysis funds, allowing model selection to be a more straightforward and less "risky" process.
- > The needs for additional research can be identified clearly, with such efforts more directed in purpose.

B. PERFORMANCE STANDARDS: SOME PRACTICAL CONSIDERATIONS

Before continuing, we point out several practical considerations that can have a direct impact on model verification. Among the most important of these are the following: data limitations (due to its form, quantity, quality, and availability); time/resource constraints; and variability in the level and timing of analysis requirements. We discuss each of these in turn.

1. Data Limitations

For a modeling simulation to be conducted, data must be gathered characterizing both the "driving forces" (emissions, meteorology, and vertical temperature profile, for example) and the "resulting effects" (pollutant concentrations). To do so requires an extensive and coordinated effort. Consequently, complete data sets usually are assembled for only a few sample days. The dates on which these data are gathered are chosen as ones likely to be typical of "worst" episode conditions. However, unanticipated shifts in meteorology (frontal passage, for example) can occur, confounding attempts to measure ambient conditions on high-concentration days. Consequently, the data available for model verification may not be representative of conditions on the day when the "second highest" concentration occurs, i.e., the worst NAAQS violation.

Confronted with such a situation, the modeler must decide the following: Even if model performance proves acceptable for non-episode conditions, can it be considered "verified" as a predictive tool for higher-concentration days? This question is part of a still more general one: Should a model be verified for more than one day, each of these days experiencing a different peak concentration? If such a procedure were followed, model performance could be evaluated for concentrations ranging from the current peak value to ones nearer the NAAQS. But, the meteorology occurring on days experiencing low peak concentrations is not typical of that occurring on high peak days. Should not the model, when used as a predictive tool, employ maximum-episode meteorology? We do not answer these questions here but note their importance as questions remaining to be resolved. We observe, however, that limitations on data quantity and availability can constrain us, limiting our flexibility in dealing with these questions.

Another difficulty can arise because of spatial limitations in the data. As we noted in the last chapter, measurement networks provide concentration data only at a few fixed sites. In general, these networks cannot guarantee observation of the "true" peak, nor are they sufficiently well-spaced to assure that the "true" concentration field can be reconstructed from the station measurements. As a practical matter, however, these station data must form the basis for the comparison of prediction and observation. Station-type performance measures, as defined in Chapter V, therefore must be the "preferred" (or rather the "unavoidable") measures of interest. We detail some of these later in Section D.

2. Time/Resource Constraints

Both the amount and quality of the data collected as well as the level of modeling analysis performed are all strongly influenced by time deadlines and resource constraints. This has several consequences among which are the following: Because it is difficult, expensive and time comsuming to mount special data gathering efforts, heavy reliance is placed on previously gathered data, even with its recognized deficiencies; also, model selection occasionally is made more on the basis of the form and extent of existing data and financial budgetary considerations than on grounds more technically justifiable. In such cases a conscious choice has been made, trading model performance for other considerations.

The combined effect of inadequate data and inappropriate model choice can reduce in value any assessment of model performance. In this report, however, we take the following view: The level of performance required of a model is determined not by exogeneous considerations but by the nature of the issue and the specific modeling application.

3. Variability of Analysis Requirements

Modeling analysis requirements differ from one application to another. There is an important question to ask in every modeling situation: How much analysis is justified? In the Los Angeles Basin, for instance, attainment of the NAAQS for ozone cannot be achieved without widespread and extensive hydrocarbon (HC) emissions control. Ambient HC levels are currently so high that more HC radicals are available than are "needed" by the chain of photochemical reactions that results in the O_3 peak. Consequently, reductions in HC emissions must be sizable before any appreciable reduction in peak O_3 can be achieved. The result of this is the following: Estimates of the percentage HC emissions control required to reach NAAQS compliance in Los Angeles are so high (75 to 80 percent) that they are not strongly sensitive to uncertainties in the value of the O_3 peak, either measured or predicted.

If the only questions to be answered depended on the general region-wide level of HC emissions control required (a SIP/C-related problem), then a fair amount of uncertainty could be tolerated in model predictions of the O_3 peak. Use of a less sophisticated model might be acceptable. Were a different issue/question addressed, however, a model providing more detailed predictions might be required.

C. MODEL PERFORMANCE ATTRIBUTES

Model predictions are subject to a number of sources of uncertainty. Some of these are data related, while others are inherent in the model theoretical formulations. Regardless of their source, however, errors manifest themselves in similar ways. They may affect a model's ability to predict peak concentrations, as well as introduce systematic bias or gross error into its predictions. They may limit a model's ability to reproduce temporal variation or affect the spatial distribution of the concentration field.

What are the attributes of desirable model performance? Ideally, we would ask that a model have five major attributes, the strength of our insistence depending on the circumstance of our application and the pollutant we are considering. The five model performance attributes are: accuracy of the peak prediction, systematic bias, lack of gross error, temporal correlation, and spatial alignment. The first of these concerns the model's ability to predict accurately the level, timing, and location of the concentration peak. The second attribute is the absence of systematic bias, where predictions are shown not to differ from observations in any consistent and unexplained way. The third attribute concerns the lack of gross error, or rather the absolute amount by which predictions differ from observations.

We classify the difference between bias and error by means of the following example. Suppose when we compare a set of model predictions with station observations, we find several large positive residuals (predicted minus observed concentrations) balanced by several equally large negative residuals. If we were testing for bias, we would allow the oppositely signed residuals to cancel. A conclusion that the model displayed no systematic bias therefore might be a justifiable one. On the other hand, were we testing for gross error, the signs of the residuals would not be considered, with oppositely signed residuals no longer allowed to cancel. Because the absolute value of the residuals is large in our example, we might well conclude that the model predictions are subject to significant gross error.

The fourth of the desirable performance attributes is that of temporal correlation. When this is important, can the model reproduce the temporal variation displayed by the observational data? A model might be judged as being capable of doing so if its predictions varied in phase with observation, that is, if they were "correlated." The fifth desirable attribute is that of spatial alignment. At each time of interest, does the model predict a concentration field that is distributed spatially like the observed one? To determine this, correlation of prediction with observation could be assessed at several points in the concentration field, e.g., monitoring stations.

The five performance attributes are interrelated. Suppose, for instance, that our model does not reproduce well the photochemistry of ozone formation in the atmosphere. Not only could its estimates of the concentration peak be in error, but also its temporal correlation and spatial alignment might be poor. Even if the model predicted the peak properly, problems might still exist. If the chemistry were "fast," the peak, though correct, might be predicted to occur sooner than that actually observed. Even if atmospheric transport were properly modeled, performance measures might then "detect" temporal and spatial problems.

By treating each performance attribute separately, we may run the risk of rejecting a model on several grounds where only a single reason actually exists. For example, slight errors in the wind field input to the model might result in predictions apparently wrong both spatially and temporally. Yet, only a single defect exists, in this case not due to the model at all.

Nevertheless, we adopt a conservative viewpoint. We suggest evaluating the model separately for the presence of each attribute, even though they themselves may be interrelated. Redundancy should not result in a satisfactory model being unfairly rejected. If model predictions are good, they will be acceptable both spatially and temporally. If they are poor, they will probably be rejected, both for temporal and spatial reasons.

If model performance is mixed, showing, for example, good temporal correlation but poor spatial alignment, two possibilities exist. Either the model performance may not be particularly poor or the performance measure used to detect one or the other performance attribute is deficient (too stringent or too lenient). In either case, however, forcing model performance to be reassessed makes sense. On balance, while requiring a model to "jump the hoop" twice may be redundant in looking for the same problem, it should provide us a measure of safety in the "double-check" it provides, presuming each attribute assumes the same importance (see the discussion below).

Although they are interrelated, the five model performance attributes are distinct. Consequently, we must employ different kinds of performance measures to determine the presence of each attribute. While we defer to Section D a statement of specific measures we recommend using, we list in Table VI-1 their objectives.

We have identified five model performance attributes. Which of these, however, is most important? This question has no unique answer, the relative importance in each problem depending on the type of issue the model is being used to address and the type of pollutant under consideration. In order to relate attribute importance to application issue in a more convenient manner, we present in Table VI-2 a matrix of generic issue class (as defined earlier in this report) and problem type. For each combination

TABLES VI-1. PERFORMANCE MEASURE OBJECTIVES

Performance Attributes	Objective of Performance Measures
Accuracy of the peak prediction	Assess the model's ability to predict the concentration peak (its level, timing and location)
Absence of systematic bias	Reveal any systematic bias in model predictions
Lack of gross error	Characterize the error in model predictions both at specific monitoring stations and overall
Temporal correlation	Determine differences between predicted and observed temporal behavior
Spatial alignment	Uncover spatial misalignment between the predicted and observed concentration fields

TABLE VI-2. IMPORTANCE OF PERFORMANCE ATTRIBUTES BY ISSUE

	Importance of Performance Attribute*						
Performance Attribute	SIP/C	AQMP	PSD	NSR	OSR	EIS/R	LIT
Accuracy of the peak prediction	1	1	1	1	2	1	1
Absence of systematic bias	1	1	1	1	1	1	1
Lack of gross error	1	1	1	1	1	1	1
Temporal correlation	2	2	3	3	3	3	3
Spatial alignment	2	2	1	3	3	3	3

---

* Category 1 - Performance standard must always be satisfied.
Category 2 - Performance standard should be satisfied, but some leeway may be allowed at the discretion of a reviewer.
Category 3 - Meeting the performance standard is desirable but failure is not sufficient to reject the model; measures dealing with this problem should be regarded as "informational."

we indicate an "importance category." We define the three categories based upon how strongly we insist our model demonstrate the presence of a given attribute. For Category 1, we require that performance standards always be satisfied (the problem type is of prime importance). For Category 2, we state that the standard should be satisfied but some leeway ought to be allowed, perhaps at the discretion of a reviewer (while the problem type is of considerable importance, some degree of "mismatch" may be tolerable). For Category 3, we are not insistent that standards be met, though we state that as being a desirable objective (the problem type is not of central importance).

A number of assumptions are embedded in Table VI-2. Among the more significant are the following:

> Both peak and "far-field" concentrations are of interest in considering PSD and NSR questions.

> Specific-source issues (PSD, NSR, OSR, EIS/R and LIT) most often deal with sources assumed to be continuously emitting at a constant level (or nearly so); consequently, performance measures considering time variations between prediction and observation are not the principal measures of interest.

> Spatial agreement between prediction and observation is particularly important in applications where PSD is an issue; this is so because source impact on pristine areas (Class I) and elevated terrain (Class II) often occurs well downwind of the source, with the magnitude and incidence of impact highly directional and spatially dependent.

> Specific-source impact generally occurs in a narrow downwind plume; thus, the monitoring network set up to provide measurement data often consists of only a few stations; as a result, the calculation of all-station performance measures may not prove meaningful.

> Error is less important in considering regional issues than is the presence of a systematic bias.

> To achieve and maintain compliance with the NAAQS (SIP/C, AQMP), alternate control strategies must be developed and evaluated. For this to be done properly, some degree of spatial resolution should be attained by the model and verified.

The relative importance of each performance attribute is dependent on the type of pollutant being considered and the averaging time required by the NAAQS. If a species is subject to a short-term standard, for instance, accuracy of the peak prediction and temporal correlation might be of considerable concern, depending on the issue being addressed. However, if the species is subject to a long-term standard, neither of these problem types are of appropriate form. We indicate in Table VI-3 a matrix of the performance attributes and pollutant species. We rank each combination by the same importance categories we used earlier in Table VI-2.

Conceivably, a conflict might exist between the ranking indicated by the issue and the pollutant matrices in Tables VI-2 and VI-3. We suggest resolving the conflict in favor of the less stringent of the two rankings. For example, suppose the issue being addressed was SIP/C and pollutant being considered was CO. According to Table VI-2, the accuracy of the peak prediction should be regarded as Category 1 (the standard must always be satisfied). However, according to Table VI-3, it should be considered as Category 2 (the standard should be satisfied but some leeway may be allowed). The conflict should be resolved by allowing the combined issue/pollutant ranking to be Category 2.

D. RECOMMENDED MEASURES AND STANDARDS

In this section we reach a major goal of this report: We identify a recommended set of performance measures and propose rationales for setting standards for each. Our discussion in this section unfolds as follows. First, we isolate a candidate list of performance measures from which we select the recommended set. Then, we detail several rationales on which to base standards for our "preferred" measures. Using these we identify specific "guiding principles" from which standards may be set. In a final

TABLE VI-3. IMPORTANCE OF PERFORMANCE ATTRIBUTES BY POLLUTANT AND AVERAGING TIME

	Importance of Performance Attribute*										
	Pollutants with Short-term Standards								Pollutants with Long-term Standards		
Performance Attribute	O_3** (1 hour)§	CO** (1 hour)	NMHC* (3 hour)	SO_2 (3 hour)	NO_2 (?)†	CO (8 hour)	TSP** (24 hour)	SO_2** (24 hour)	NO_2** (1 year)	TSP (1 year)	SO_2 (1 year)
Accuracy of the peak prediction	1	1	1	1	1	1	1	1	3	3	3
Absence of systematic bias	1	1	1	1	1	1	1	1	1	1	1
Lack of gross error	1	1	1	1	1	1	1	1	1	1	1
Temporal correlation	1	2	2	2	1	2	3	3	N/A††	N/A	N/A
Spatial alignment	1	2	2	2	1	2	2	2	2	2	2

---

* Category 1 - Performance standard must be satisfied.
Category 2 - Performance standard should be satisfied, but some leeway may be allowed at the discretion of a reviewer.
Category 3 - Meeting the performance standard is desirable but failure is not sufficient to reject the model.

† No short-term NO_2 standard currently exists.

§ Averaging times required by the NAAQS are in parentheses.

** Primary standards.

†† The performance attribute is not applicable.

synthesis, we present a summary table listing for each performance attribute, the recommended measures and a means for setting standards for them, along with a sample value for the standard (ones listed are appropriate for the Denver case study described in Section E of this chapter).

1. Recommended Performance Measures

Of the many performance measures considered in Chapter V (and in more detail in Appendix C), which of these are most suitable for use in establishing standards for model performance? The answer to this is constrained in two major ways, the first conceptual and the second practical. First, the conceptual constraint is imposed by the types of performance attributes we are concerned with: The measures must adequately assess the presence or absence of each of the five attributes. Second, the practical constraint is imposed by the "sparseness" of the observational data: Since station observations constitute the only data available for characterizing "true" ambient conditions, we have little choice but to employ station performance measures in determining model acceptability.

We draw a distinction between those measures that are of general use in examining model performance and the much smaller subset of them that is most amenable to the establishment of explicit standards. Many measures can provide rich insight into model behavior but the information is conveyed in a qualitative way not suitable for quantitative characterization (a requisite for use in setting performance standards). These "measures," often involving graphical display, really are tools for use in "pattern recognition." They display model behavior in suggestive ways, highlighting "patterns" whose presence reveals much about model performance. Several examples of such "measures" are isopleth contour maps of predicted concentrations and estimated "observed" ones, isopleth contour maps of the differences between the two, and time histories of predicted and observed concentrations at specific monitoring stations.

Though we focus on station measures for use in setting model performance standards, we do not suggest the calculation of performance measures be limited to them. Many others, where each is appropriate, should be used. The data should be viewed in as many, varied ways as possible in order to enrich insight into model behavior. We suggest a number of useful measures both in Chapter V and Appendix C.

Given that station measures are our "preferred" (rather, our "unavoidable") choice, we now consider the list of candidate measures. From these we select our final recommended set. We present the candidate station performance measures in Table VI-4. We group them by the number of stations compared noting the performance attribute and generic issue class they are most suited for addressing. We identify four types of comparisons:

> Event Specific Values. Predicted and observed concentrations are compared at the time a specific event occurs. For instance, the peak station prediction can be compared with the peak station observation, even though these may occur at different stations and times.

> Comparative Values. Predicted and observed concentrations are compared at the same monitoring station.

> Average Values. Predicted and observed concentrations are compared averaged for all monitoring stations.

> Offset Values. Observed concentrations at a given station are compared with predicted values offset by a small amount spatially (values at near-by stations) and/or temporally (values at other times, either earlier or later).

Performance measures are of two different kinds: "absolute" and "informational." The first type includes those measures for which we can set specific, absolute standards. Measures of the second type are more informational in nature, providing qualitative insight into model performance. Their values are to be considered as "advisory," having associated with them no specific standard.

TABLE VI-4. CANDIDATE STATION PERFORMANCE MEASURES

				Issue Category						
		Performance Measure		Multiple-Source		Specific-Source				
Stations Considered	Performance Attributes	Description	Status	SIP/C	AQMP	PSD	NSR	OSR*	EIS/R	LIT
Peak Stations (Event-Specific Values)	Accuracy of the peak prediction (Concentration level)	1. Difference between or ratio of peak station concentrations (could be at different measurement stations)	Absolute	X	X	X	X	X	X	X
		2. Difference between or ratio of predicted and observed concentrations at the station recording the maximum measured value	Absolute	X	X		X		X	X
	Accuracy of the peak prediction (Location of Peak)	3. Spatial displacement between predicted and observed peak stations	Informational	X	X		X		X	X
	Accuracy of the peak prediction (Timing of Peak)	4. Timing difference between occurrence of predicted and observed peak	Absolute	X	X					
Each Station Separately (Comparative Values)	Absence of systematic bias	5. Average relative deviation	Absolute	X	X	X	X	X	X	X
	Lack of gross error	6. Average absolute relative deviation	Absolute	X	X	X	X	X	X	X
		7. Standard deviation of deviations	Absolute	X	X	X	X	X	X	X
	Temporal correlation/spatial alignment	8. Correlation coefficient	Absolute	X	X					
	Temporal correlation	9. Temporal offset correlation coefficient	Informational	X	X					
		10. Plots of comparative time histories	Informational	X	X					
All Stations Together (Average Values)	Absence of systematic bias	11. Average relative deviation	Absolute	X	X	X	X	X	X	X
	Lack of gross error	12. Average absolute relative deviation	Absolute	X	X	X	X	X	X	X
		13. Standard deviation of deviations	Absolute	X	X	X	X	X	X	X
		14. Correlogram of prediction-observation pairs	Informational	X	X					
		15. Ratio of peak to average deviation	Informational	X	X					
	Temporal correlation/spatial alignment	16. Correlation coefficient	Absolute	X	X					

TABLE VI-4 (Concluded)

					Issue Category						
		Performance Measure			Multiple-Source		Specific-Source				
Stations Considered	Problem Type		Description	Status	SIP/C	AQMP	PSD	NSR	OSR*	EIS/R	LIT
	Temporal correlation	17.	Temporal offset correlation coefficient	Informational	X	X					
		18.	Plot of comparative time histories	Informational	X	X					
Nearby Stations (Offset Values)	Spatial alignment	19.	Spatial offset correlation coefficient (comparison at the same time)	Informational	X	X					
	Temporal correlation/ Spatial Alignment	20.	Spatial/temporal offset correlation coefficient (comparison at different times)	Informational	X	X					

---

* These measures are appropriate if offsets are considered at the level of ambient concentrations rather than primary emissions.

Often in practice modeling predictions are known with greater spatial resolution than measurement data. The predicted concentration field, for instance, can be resolved at intervals of several kilometers or less by various types of models, including grid and Gaussian ones. To retain the information contained in concentration field predictions, several "hybrid" performance measures can be employed. With these, concentration field predictions are compared with station measurements. We list in Table VI-5 several of these hybrid measures. When predictions are available in this more detailed form, these measures may be calculated to supplement those in Table VI-4.

Our recommended choice of performance measures is based upon the following criteria:

- The measure is an accurate indicator of the presence of a given performance attribute.
- The measure is of the "absolute" kind, that is, specific standards can be set.
- Only station measures should be considered for use in setting standards. (This is more an unavoidable choice than a preferred one.)

Based on these criteria, we have selected the set of measures described in Table VI-6. The use of ratios (C_{p_p}/C_{p_m} and $\bar{\mu}$, for example) can introduce difficulties: They can become unstable at low concentrations, and the statistics of a ratio of two random variables can become troublesome. Nevertheless, when used properly, their advantages can be offsetting. For example, the use of C_{p_p}/C_{p_m} instead of ($C_{p_p}-C_{p_m}$) permits a health effects rationale to be used in recommending a performance standard (see a later discussion of the effects rationale).

Before continuing, however, we insert an important caveat. For calculation of these measures to be statistically meaningful, a certain minimum level of spatial and temporal "richness" must be available from monitoring data. Often, this criterion is met for multiple-source, urban applications. However, for isolated point source applications, it may not be. For such cases, data inadequacies may be overcome by using prototypical "test bed" data bases for the purposes of model verification. Selection of the proper "test bed" could be accomplished by choosing the prototypical data base that describes an application most nearly like the proposed one.

These data bases, where they do not already exist, could be assembled through special measurement efforts at existing large point sources. Monitoring could be extensive enough to insure adequate data "richness."

As a practical matter, however, such "test beds" are not currently available. Verification instead must be conducted using whatever data are at hand. These may be provided by tracer experiments. Alternatively, where a source already exists (for instance, where retrofit of pollution control equipment is the issue or where construction of a new source is to occur on the site of an existing source), some site-specific data already may be available.

Considerable care should be exercised when using such data to calculate the performance measures listed in Table VI-6. If the data are too "sparse," in either a spatial or a temporal sense, these measures may be of little value, or worse yet, may actually be misleading. Additional work needs to be conducted to identify, if possible, supplementary performance measures for use when the available data is inadequate for reliable use of the recommended measures.

Having stated the above caveat, we continue. A number of key assumptions are embedded in the choice of the specific measures shown in Table VI-5. We state several of them:

> Concentration gradients within a pollutant cloud can be "steep". Thus a slight spatial misalignment of the cloud, perhaps an unconsequential problem on its own, can sometimes result in the predicted peak occurring at a different monitoring station than the measured peak. Estimating the value of the concentration peak, however, is often of much greater importance than predicting its exact location.

TABLE VI-5. USEFUL HYBRID PERFORMANCE MEASURES

Stations Considered	Performance Attribute	Performance Measure			Issue Category						
					Multiple-Source		Specific-Source				
		Description		Status	SIP/C	AQMP	PSD	NSR	OSR*	EIS/R	LIT
Peak Station (Event-Specific Values)	Accuracy of the peak prediction (Concentration level)	1.	Difference between or ratio of predicted peak concentration and highest station value	Absolute	X	X	X	X	X	X	X
	Accuracy of the peak prediction (Location of Peak)	2.	Spatial displacement between the predicted peak and the station measuring the highest value	Informational	X	X		X		X	X
	Accuracy of the peak prediction (Timing of Peak)	3.	Timing difference between occurrence of the predicted peak and the maximum station measurement	Informational	X	X					
Each Station Separately (Comparative Values)	Spatial alignment	4.	Plot showing for each hour during the day the distance and direction from the measurement station to the nearest point at which a predicted concentration occurs equal to the station measured value	Informational	X	X	X				
All Stations Together (Average Values)	Lack of gross error	5.	Difference for each hour between the average predicted concentration (averaged over the entire field) and the average station measurement (averaged over all stations)	Informational	X	X	X	X	X	X	X
		6.	Difference for each hour between the standard deviations of the predicted concentrations and the station measured values	Informational	X	X	X	X	X	X	X

TABLE VI-6 . MEASURES RECOMMENDED FOR USE IN SETTING MODEL PERFORMANCE STANDARDS†

Performance Attribute	Performance Measure
Accuracy of the peak prediction	Ratio of the predicted station peak to the measured station (could be at different stations and times) C_{p_p}/C_{p_m} Difference in timing of occurrence of station peak* Δt_p
Absence of systematic bias	Average value and standard deviation of the mean deviation about the perfect correlation line normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either the predicted or the observed values exceed some appropriate minimum value (possibly the NAAQS) $(\overline{\mu}, \sigma_{\overline{\mu}})_{\mathrm{OVERALL}}$
Lack of gross error	Average value and standard deviation of the absolute deviation about the perfect correlation line normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either the predicted or the observed values exceed some appropriate minimum value (possibly the NAAQS) $\left(\lvert\overline{\mu}\rvert, \sigma_{\lvert\overline{\mu}\rvert}\right)_{\mathrm{OVERALL}}$
Temporal correlation*	Temporal correlation coefficients at each monitoring station for the entire modeling period and an overall coefficient averaged for all stations $r_{t_i}, r_{t_{\mathrm{OVERALL}}}$, for $1 \le i \le M$ monitoring stations
Spatial alignment	Spatial correlation coefficients calculated for each modeling hour considering all monitoring stations, as well as an overall coefficient average for the entire day $r_{x_j}, r_{x_{\mathrm{OVERALL}}}$ for $1 \le j \le N$ modeling hours

---

* These measures are appropriate when the chosen model is used to consider questions involving photochemically reactive pollutants subject to short-term standards.

† There is deliberate redundancy in the performance measures. For example, in testing for systematic bias, $\overline{\mu}$ and $\sigma_{\overline{\mu}}$ are calculated. The latter quantity is a measure of "scatter" about the perfect correlation line. This is also and indicator of gross error and could be used in conjunction with $\lvert\overline{\mu}\rvert$ and $\sigma_{\overline{\mu}}$.

Consequently, we suggest, when this seems reasonable (judgment is necessary here), comparing the peak station prediction with the peak station measurement, regardless of when or where they both occur.

> In addressing questions involving pollutants subject to short-term standards, diurnal variation occurs in concentration levels. It is reasonable to insist short-term predictions emulate that pattern. Differences in the timing of the peak should be considered (particularly for photochemically reactive pollutants) and temporal correlation should be evaluated.

> In many circumstances, percentage differences between predicted and observed concentrations seem better indicators of model performance than gross differences. For instance, a difference of 0.04 ppm of ozone might be regarded as serious if ambient levels were 0.10 ppm where it might not be if those levels were 0.24 ppm. The use of such measures can cause some problems: Ratios can become unstable at low concentrations, and the statistics of a ratio of two random variables can be complex. Nevertheless, percentage differences should be calculated (possibly along with gross differences). Further, we suggest that residuals (prediction minus observation) be taken about the perfect correlation line (prediction equals observation), since we have no *a priori* reason to regard observation as any more accurate than prediction. This was pointed out by Anderson et al. (1977). We also suggest normalizing the residuals by the arithmetic average of the predicted and observed concentration.

> The concentrations of greatest interest are often the higher values, that is, those that exceed some appropriate minimum value (possibly the NAAQS, though this may differ from one situation to another). We may be less interested in model reliability below those levels. We suggest that performance measures include only those prediction-observation "pairs" where one or the other value exceeds the chosen minimum value. (Possibly "stratification" may be of interest, that is, repeating the calculation of measures using different minimum values).

This should not be done, however, if it results in the number of pairs being reduced below the number required for statistical significance.

- Measurement stations usually are widely spaced. We assumed this spacing to be so great that the use of spatial/temporal offset correlation coefficients would be of uncertain value. Consequently, we did not include them among the list of measures recommended for use.
- Redundancy should be built into the calculation of performance measures. This provides an internal means for double-checking results. For example, in testing for systematic bias, $\overline{\mu}$ and $\sigma_{\overline{\mu}}$ are calculated. The latter quantity is a measure of "scatter" about the perfect correlation line. This is also an indicator of gross error and should be used in conjunction with $|\overline{\mu}|$ and $\sigma_{|\overline{\mu}|}$.

2. Recommended Performance Standards

Having identified the performance measures requiring a specific standard, we now consider four alternative rationales for setting those standards. We designate the four as follows:

- Health Effects
- Control Level Uncertainty
- Guaranteed Compliance
- Pragmatic/Historic

The guiding principles for each of these rationales are stated in Table VI-7.

We describe in detail each rationale in Appendix D, deferring their technical description in order not to interrupt the flow of this chapter. However, to offer insight into their general nature, we present here a brief outline of each.

TABLE VI-7. POSSIBLE RATIONALES FOR SETTING MODEL PERFORMANCE STANDARDS

Rationale	Guiding Principle
Health Effects	The metric of concern is the area-integrated cumulative health effects due to pollutant exposure; the ratio of the metric's value based on prediction to its value based on observation must be kept to within a prescribed tolerance of unity.
Control Level Uncertainty	Uncertainty in the percentage of emissions control required must be kept within certain allowable bounds.
Guaranteed Compliance	Compliance with the NAAQS must be "guaranteed;" all uncertainty must be on the conservative side even if its means introducing a systematic bias.
Pragmatic/Historic	In each new application of a model should perform at least as well as the "best" previous performance of a model in its generic class in a similar application; until such a historical data base is complete, other more heuristic approaches may be applied.

> Health Effects. The most fundamental reason for setting air quality standards is to limit the adverse health impact the regulated pollutants (and their products) produce. Thus, founding a model performance standard on a health effects basis has strong intuitive appeal. To do so, we assume an analytic form for urban population distribution and an exposure/dosage health effects functional, both of which require as inputs only easily derived data. Using these, we determine in analytic form a new health-based metric: the area-integrated cumulative health effects. We estimate through this metric the total health burden experienced by the population during the day. The model is required to predict concentrations that do not differ from observations to the point an unacceptable difference is seen in the health metric. While the data used is application-specific, the method itself is general. The assumptions made in deriving

this rationale, while extensive, seem plausible. A sample case was conducted for ozone exposure in the Denver Metropolitan region, with promising corroboration of the rationale in several key regards. The sample case is described in detail in Appendix D.

> Control Level Uncertainty. With this rationale we set performance standards to ensure that uncertainty in estimates of the amount of pollution control required be kept within acceptable bounds. These limits may be determined in a number of ways, but we consider limits on uncertainty in control cost as a promising means for doing so. If we can assume that pollutant production and evolution over the modeled region can be approximated by some simple surrogate, such as an isopleth diagram for ozone, then control uncertainty limits can be directly and easily related to equivalent bounds in uncertainty in the pollutant peak, the quantity to which control strategies are often designed.

> Guaranteed Compliance. The NAAQS are written in quite specific terms and must ultimately be complied with. An argument can be made that to "guarantee" such compliance, uncertainty in model predictions must be on the "conservative" side. That is, the probability must be acceptably small that a control strategy designed based on model predictions will not actually achieve compliance. We consider this rationale here and in Appendix D primarily for completeness. While the rationale has some potential usefulness, it implies the introduction of a systematic bias into modeling results, something we would hope to avoid in a final choice of a performance standard.

> Pragmatic/Historic. Standards for all performance measures cannot be derived based on the rationales mentioned above, something we will discuss later in this chapter. Until additional research expands our options by providing insight into other rationales, we adopt a pragmatic approach. We may proceed in either of two ways. If we are able to state

heuristically a specific guiding principle for setting a standard for a particular measure, we invoke it. Otherwise, we simply require the following: In each new application a model should perform at least as well as the "best" previous performance of a model in its generic class in a similar application. In addition to being pragmatic, this last approach is also evolutionary, requiring a continually expanding and updated model/application data base.

The four rationales differ in their usefulness vis-a-vis the five performance attributes. Shown in Table VI-8 are the attributes addressable by measures whose standards are set by each of the rationales. Only the Pragmatic/Historic rationale is of use in addressing all attributes; the other three are of use principally in defining the level of performance required in predicting values at or near the concentration peak. The Health Effects and Guaranteed Compliance rationales also may have some application to problems involving concentration field error.

TABLE VI-8. PERFORMANCE ATTRIBUTES ADDRESSABLE USING PERFORMANCE STANDARD RATIONALES

Performance Attribute	Health* Effects	Control Level* Uncertainty	Guaranteed Compliance	Pragmatic/ Historic
Accuracy of the peak prediction	X	X	X	X
Absence of systematic bias				X
Lack of gross error	X		X	X
Temporal correlation				X
Spatial alignment				X

---

* These are most suited for photochemically reactive pollutants subject to short-term standards.

One conclusion seems clear. Unless more comprehensive rationales are developed in subsequent research work, several must be used simultaneously to completely define standards of performance. Any one of the four can be used to specify allowable bounds on model performance in predicting peak concentrations. Either the Health Effects or the Pragmatic/Historic rationales can be helpful in setting standards for error measures. Only the latter of these two rationales is of use for addressing attributes of the other types.

We associate in Table VI-9 each rationale with those generic issues for which its use is appropriate. Several assumptions are embedded in that table. Among them are the following:

> Health effects are not of overriding concern in PSD and OSR issues, for reasons noted earlier. (Even though we indicate such a rationale may be used in addressing other specific-source issues, we observe that plume "narrowness" can limit downwind health impact).

> Near-source peak concentrations are not of primary interest in OSR, but rather "far-field" average values.

> The Guaranteed Compliance rationale is of use in addressing questions involving PSD as long as the air quality standards being used are the PSD class increments.

TABLE VI-9. ASSOCIATION OF RATIONALES WITH GENERIC ISSUES

	Issue Category						
	Multiple-Source		Specific-Source				
Rationale	SIP/C	AQMP	PSD	NSR	OSR	EIS/R	LIT
Health Effects	X	X	X	X		X	X
Control Level Uncertainty	X	X	X	X		X	X
Guaranteed Compliance	X	X	X	X		X	X
Pragmatic/ Historic	X	X	X	X	X	X	X

Having outlined the rationales we consider in this report, it remains to match them with the set of performance measures we recommended earlier in this chapter. As is clear from Table VI-8, we have no alternative but to apply the Pragmatic/Historic rationale for those measures designed to test for systematic bias or to evaluate temporal behavior and spatial alignment. However, several alternatives exist for measures dealing with peak performance and gross error.

We select in the following ways from among the alternatives. Hoping to avoid introducing a procedural bias, we first eliminate the Guaranteed Compliance rationale from further consideration. Then, because the Health Effects rationale is better suited for use in setting standards for peak-accuracy measures, we choose to use it only in that way.

Our recommended choice for use in establishing standards for peak-accuracy measures is a composite one, combining the Health Effects and Control Level Uncertainty rationales. Were a model to overpredict the peak, a control strategy designed based on its prediction might be expected to abate the health impact actually occurring. If the model underpredicted, however, the control strategy might be "underdesigned," with the risk existing that some of the health impact might remain unabated even after control implementation. The penalty, in a health sense, is incurred only when the model underpredicts. The Health Effects rationale then is one-sided, helping us set performance standards only on the "low side."

On the other hand, the Control Level Uncertainty rationale is bounded "above" and "below", that is, its use provides a tolerance interval about the value of the measured peak concentration. For a model to be judged acceptable under this criterion, its prediction of the peak concentration would have to fall within this interval. Model underprediction could lead to control levels lower than required, but residual health risks. Overprediction, on the other hand, could lead to abatement strategies posing little or no health risk but incurring control costs greater than required.

For the above reasons, we suggest that the Control Level Uncertainty rationale be used to establish an upper bound (overprediction) on the acceptable difference between the predicted and observed peak. We would choose the lower bound (underprediction) to be the interval that is the minimum of that suggested by the Health Effects and Control Level Uncertainty rationales.

We list our recommendations in Table VI-10, noting the possibility for peak-accuracy measures that the recommended rationales may not be appropriate in all applications for all pollutants. Whether health effects would be an appropriate consideration when considering TSP, for instance, is unclear. The Health Effects rationale is best suited for use in urban applications involving short-term, reactive pollutants. In those circumstances when the HE or CLU rationales are not suitable, we suggest the Pragmatic/Historic rationale.

TABLE VI-10. RECOMMENDED RATIONALES FOR SETTING STANDARDS

Performance Attribute	Recommended Rationale
Accuracy of peak prediction	Health Effects* (lower side/underprediction) Control Level Uncertainty* (upper side/overprediction)
Absence of sytematic bias	Pragmatic/Historic
Lack of gross error	Pragmatic/Historic
Temporal correlation	Pragmatic/Historic
Spatial alignment	Pragmatic/Historic

---

* These may not be appropriate for all regulated pollutants in all applications. When they are not, the Pragmatic/Historic rationale should be employed. They are most applicable for photochemically reactive pollutants subject to a short-term standard (O_3 and NO_2, if a 1-hour standard is set).

3. Summary Table of Recommended Measures and Standards

Until now, our discussion has remained general when relating performance measures and standards. Here we become specific. In Table VI-11, we summarize for each of the five problem types whose presence we are testing for the performance measures we recommend and the standards we suggest. Since the actual value of the standard may vary from one application to another or between pollutant types, we present sample values calculated based on a sample case. The example is appropriate for consideration of SIP/C in the Denver Metropolitan region and is described in a case study fashion in Section E of this chapter.

Where we invoke the Pragmatic/Historic rationale as justification for selecting specific standards, we also state the specific guiding principle we followed. We summarize those here:

> When the pollutant being considered is subject to a short-term standard, the timing of the concentration peak may be an important quantity for a model to predict. This is particularly true when the pollutant is also photochemically reactive. We state as a guiding principle: "For photochemically reactive pollutants, the model must reproduce reasonably well the phasing of the peak." For ozone an acceptable tolerance for peak timing might be ± 1 hour.

> The model should not exhibit any systematic bias at concentrations at or above some appropriate minimum value (possibly the NAAQS) greater than the maximum resulting from EPA-allowable calibration error. We would consider in our calculations any prediction-observation pair in which either of the values exceed the pollutant standard. Error (as measured by its mean and standard deviation) should be indistinguishable from the distribution of differences resulting from the comparison of an EPA-acceptable monitor with an EPA reference monitor. The EPA has set maximum allowable limits on the amount by which a monitoring technique may differ from a reference method (40 CFR §53.20). An

TABLE VI-11. SUMMARY OF RECOMMENDED PERFORMANCE MEASURES AND STANDARDS

		Performance Standard		
Performance Attribute	Performance of Measure	Type of Rationale	Guiding Principle	Sample Value (Denver Example)
Accuracy of the peak prediction	Ratio of the predicted station peak to the measured station peak (could be at different stations and times) C_{p_p}/C_{p_m}	Health Effects† (lower side) combined with Control Level Uncertainty (upper side)	Limitation on uncertainty in aggregate health impact and pollution abatement costs†	$80 \le \frac{C_{p_p}}{C_{p_m}} \le 150$ percent
	Difference in timing of occurrence of station peak* Δt_p	Pragmatic/Historic	Model must reproduce reasonably well the phasing of the peak, say, ±1 hour	± 1 hour
Absence of systematic bias	Average value and standard deviation of mean deviation about the perfect correlation line normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either predicted or observed values exceed some appropriate minimum value (possibly the NAAQS). $(\overline{u}, \sigma_{\overline{u}})_{OVERALL}$	Pragmatic/Historic	No or very little systematic bias at concentrations (predictions or observations) at or above some appropriate minimum value (possibly the NAAQS); the bias should not be worse than the maximum bias resulting from EPA-allowable monitor calibration error (-8 percent is a representative value for ozone); the standard deviation should be less than or equal to that of the difference distribution of an EPA-acceptable monitor** compared with a reference monitor. (3 pphm is representative for ozone at the 95 percent confidence level)	No apparent bias at ozone concentrations above 0.06 ppm (see Table VI-12 and Figures VI-5 and VI-6 for further details)
Lack of gross error	Average value and standard deviation of absolute mean deviation about the perfect correlation line normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either predicted or observed values exceed some appropriate minimum value (possibly the NAAQS) $(\lvert\overline{u}\rvert, \sigma_{\lvert\overline{u}\rvert})_{OVERALL}$	Pragmatic/Historic	For concentrations at or above some appropriate minimum value (possibly the NAAQS), the error (as measured by the overall values of $\lvert\overline{u}\rvert$ and $\sigma_{\lvert\overline{u}\rvert}$) should be indistinguishable from the difference resulting from comparison of an EPA-acceptable monitor with a reference monitor	NO excessive gross error (see Table VI-12 and Figures VI-5 and VI-6 for further details)
Temporal correlation*	Temporal correlation coefficients at each monitoring station for the entire modeling period and an overall coefficient for all stations $r_{t_i}, r_{t_{OVERALL}}$, for $1 \le i \le M$ monitoring stations	Pragmatic/Historic	At a 95 percent confidence level, the temporal profile of predicted and observed concentrations should appear to be in phase (in the absence of better information, a confidence interval may be converted into a minimum allowable correlation coefficient by using an appropriate t-statistic)	For each monitoring station, $0.69 \le r_{t_i} \le 0.97$ Overall, $r_{t_{OVERALL}} = 0.88$ In this example a value of $r \ge 0.53$ is significant at the 95 percent confidence level
Spatial alignment	Spatial correlation coefficients calculated for each modeling hour considering all monitoring stations, as well as an overall coefficient for the entire day $r_{x_j}, r_{x_{OVERALL}}$, for $1 \le j \le N$ modeling hours	Pragmatic/Historic	At a 95 percent confidence level, the spatial distribution of predicted and observed concentrations should appear to be correlated	For each hour, $-0.43 \le r_{x_j} \le 0.66$ Overall, $r_{x_{OVERALL}} = 0.17$ In this example a value of $r \ge 0.71$ is significant at the 95 percent confidence level

---

* These measures are appropriate when the chosen model is used to consider questions involving photochemically reactive pollutants subject to short-term standards.

† These may not be appropriate for all regulated pollutants in all applications. When they are not the Pragmatic/historic rationale should be employed.

** The EPA has set maximum allowable limits on the amount by which a monitoring technique may differ from a reference method. An "EPA-acceptable monitor" is defined here to be one that differs from a reference monitor by up to the maximum allowable amount.

"EPA-acceptable monitor" is defined here to be one that differs from a reference monitor by up to the maximum allowable amount.

> Prediction and observation should appear to be correlated at a 95 percent confidence level, both when compared temporally and spatially. We can estimate the mininum allowable value for the respective correlation coefficient by using a t-statistic at the appropriate percentage level and having the degrees of freedom required by the number of prediction-observation pairs.

The guiding principles noted above are plausible ones, though in some cases they are arbitrary. As a "verification data base" of experience is assembled, historically achieved performance levels may be better indicators of the expected level of model performance. Standards derived on this more pragmatic basis may supplant those deriving from the "guiding principles" followed in this report.

4. Formulas for Calculating Performance Measures and Standards

A number of performance measures are recommended in Table VI-6. Here we state explicitly the equations used for their calculation and the forms assumed by the standards. We include, where appropriate, brief theoretical justifications for these relationships.

The definitions are self-explanatory for measures testing the accuracy of the peak model prediction. Specifically,

$$\alpha \leq \frac{C_{P_p}}{C_{P_m}} \leq \beta \quad , \qquad \text{(VI-1)}$$

where C_{P_p} is the peak station prediction, C_{P_m} is the peak station measurement, α is the lower bound on the ratio of the peaks, and β is the upper bound. The bounds may be determined either from Pragmatic/Historic considerations or, where possible, by means of the Health Effects/Control Level Uncertainty rationales described in Appendix D. The latter of these two approaches may prove feasible only when considering photochemically reactive pollutants (particularly ozone) subject to a short-term standard. Also, for such reactive species,

$$|\Delta t_p| \leq \delta \quad , \qquad \text{(VI-2)}$$

where $|\Delta t_p|$ is the absolute value of the difference between the predicted and observed times of the station peak, and δ is the maximum allowable difference, say, one hour (this is an arbitrarily set value).

Underlying our definitions of bias and error is the following assumption: A priori, we have no reason to prefer either prediction or observation as a better measure of reality. Both, in fact, can be subject to significant uncertainty. It follows from this assumption that residuals (predicted concentrations minus observed ones) should be taken perpendicularly about the perfect correlation line.

We emphasize an important point: The residual for a given prediction-observation pair is not the geometric distance from the perfect correlation line, as displayed in a correlogram (such as the one shown later in Figure VI-3). Rather, the geometric distance must be scaled downward by a factor of $\sqrt{2}$. That this is so follows from the discussion presented below. It is based on our requirement that prediction and observation differ by no more than the maximum amount by which an EPA-acceptable monitoring technique may differ from the accepted reference technique.

Uncertainty in monitoring results can be introduced from many sources. Three principle source categories are the calibration method, the agreement with the reference monitoring technique, and the actual instrument error. The last of these categories includes instrument noise and precision, measurement drift, and interference from other contaminants. In defining the characteristics of the EPA-acceptable monitor we wish to use as a standard, we have chosen to include only the first two error source categories. We thus eliminate the need to consider performance characteristics of specific monitoring instruments. Also, in comparing a monitor with an instrument using the EPA-accepted reference monitoring technique, it is not unreasonable to assume that both are subject to the same instrument error.

We may define an acceptance standard for a model insofar as error and bias are concerned: The distribution of differences between prediction and observation must be indistinguishable from that resulting from the comparison of an EPA-acceptable monitor with the accepted reference monitor. Specifically, we define "indistinguishable" to mean

$$\xi \leq \overline{\mu} \leq \xi \quad , \tag{VI-3}$$

$$\sigma_{\overline{\mu}} \leq \varepsilon \quad , \tag{VI-4}$$

where ξ and ε can be determined from federal regulations (40 CFR §53.20) for instrument performance, and $\overline{\mu}$ and $\sigma_{\overline{\mu}}$ are defined below.

We may confirm a model's acceptability by hypothesizing that the acceptance standard for bias and error is satisfied and checking to determine whether this hypothesis is violated. Consistent with this approach, we may assume that each prediction and observation pair are random samples drawn from the same distribution, the one that describes the behavior of an EPA-acceptable monitor with respect to a reference monitor. The standard deviation (S.D.) of a random variable whose value is the difference of two other random variables having the same S.D. σ may be expressed as

$$\sigma_D = \sqrt{2}\,\sigma \quad . \tag{VI-5}$$

The geometric distance from the perfect correlation line, d_i, may be written as

$$d_i = \frac{P_i - M_i}{\sqrt{2}} \quad , \tag{VI-6}$$

where P_i and M_i are the i-th prediction-observation pair. We are searching for a test variable σ_μ to compare with σ. Therefore, referring to Equation VI-5, we see that we must divide d_i by $\sqrt{2}$ to obtain the properly scaled mean deviation from the perfect correlation line, d_i^*, that is,

$$d_i^* = \frac{P_i - M_i}{2} \quad . \tag{VI-7}$$

Thus, the average and standard deviation of the mean deviation may be expressed as

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{P_i - M_i}{2}\right) \tag{VI-8}$$

$$\sigma_\mu = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{P_i - M_i}{2} - \mu\right)^2} \quad . \tag{VI-9}$$

These quantities may be compared with those characterizing the distribution of differences between an EPA-acceptable monitor and a reference instrument. Those values may be derived from 40 CFR §53.20. As an example, (see Burton, et al., 1976) an EPA-acceptable monitor for ozone/oxidants could have a -8 percent bias and a 95 percent confidence interval of ±3 pphm (a σ of 1.53 ppm). If an EPA-acceptable monitor were defined to be subject to instrument error as well, the -8 percent bias would remain because it is assumed due to calibration, but the 95 percent confidence interval would increase to ±7 pphm (a σ of 3.57 ppm).

We noted earlier that the "seriousness" of the magnitude of a given residual depends on the ambient concentration of the pollutant being considered. For instance, a value for d_i^* of 2 pphm might be considered of less importance when ambient concentrations are on the order of 30 pphm than when they are 10 pphm. In consideration of this effect, we suggest normalizing residuals by the arithmetic average of the predicted and observed concentrations for a given pair. This is consistent with our earlier statement that, a priori, we have no reason to prefer observation over prediction as inherently better indicators of reality.

Defining the average concentration to be

$$C_{AVE} = \frac{P_i + M_i}{2} \quad , \tag{VI-10}$$

we may write expressions for the normalized average and standard deviation of the mean deviation about the perfect correlation line:

$$\overline{\mu} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{P_i - M_i}{P_i + M_i}\right) \tag{VI-11}$$

$$\sigma_{\overline{\mu}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{P_i - M_i}{P_i + M_i} - \overline{\mu}\right)^2} \tag{VI-12}$$

A deliberate redundancy has been built into the list of suggested performance measures. Both σ_{μ} and $\sigma_{\overline{\mu}}$ are measures of "scatter" about the perfect correlation line. Thus, they are also indicators of gross error and may be used in conjunction with those measures explicitly listed in Table VI-6 for use in investigating gross error. These measures consider absolute rather than signed residuals. Specifically the normalized average value and standard deviation of the absolute deviation about the perfect correlation line may be written

$$|\overline{\mu}| = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|P_i - M_i|}{P_i + M_i}\right) \tag{VI-13}$$

$$\sigma_{|\bar{\mu}|} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{|P_i - M_i|}{P_i + M_i} - |\bar{\mu}|\right)^2} \quad . \qquad \text{(VI-14)}$$

Their values may be compared with standards such that

$$|\bar{\mu}| \leq \lambda \quad , \qquad \text{(VI-15)}$$

$$\sigma_{|\bar{\mu}|} \leq \gamma \quad , \qquad \text{(VI-16)}$$

when the values of λ and γ may be derived from instrument performance specifications in federal regulations.

It may be helpful to visualize the definitions of d_i^* and C_{AVE} geometrically on a correlogram. Figure VI-1 is a schematic, showing the orientation of the d^*-C_{AVE} axes with respect to the P-M axes of the correlogram. The C_{AVE} axis is aligned with the perfect correlation line, and both the d^* and C_{AVE} axes are scaled downward by a factor of $\sqrt{2}$ from the P and M axes.

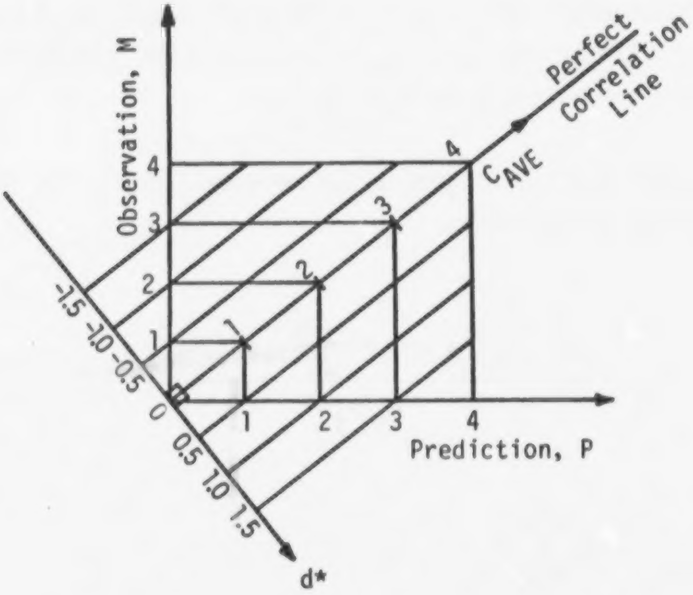

Observation, M
4
3
2
1
0
Prediction, P
1
2
3
4
Perfect Correlation Line
C_AVE
1
2
3
4
-1.5
-1.0
-0.5
0
0.5
1.0
1.5
d*

FIGURE VI-1. ORIENTATION AND SCALING OF C_{AVE} AND d^* AXES ON A PREDICTION-OBSERVATION CORRELOGRAM

Finally, we consider measures suitable for use in testing for temporal correlation and spatial alignment. The former of these is of concern when the chosen model is used to consider questions involving photochemically reactive pollutants subject to a short-term standard. We suggest the use of temporal correlation coefficients, whose values are defined to be

$$r_{t_i} = \frac{\frac{1}{N-1}\sum_{j=1}^{N}\left(P_{i,j} - \mu_{P_i}\right)\left(M_{i,j} - \mu_{M_i}\right)}{\sigma_{P_i}\sigma_{M_i}} \qquad \text{(VI-17)}$$

$$r_{t_{OVERALL}} = \frac{1}{K}\sum_{i=1}^{K} r_{t_i} \quad , \qquad \text{(VI-18)}$$

where r_{t_i} is the temporal correlation coefficient at the i-th station for the N divisions of the modeling period, and $r_{t_{OVERALL}}$ is the average correlation coefficient for all the K monitoring stations. Also, μ_{P_i} and σ_{P_i} are the mean and standard deviations of the predictions for N hours at the i-th station. Similarly, μ_{M_i} and σ_{M_i} are the mean and standard deviations of the concentrations at the i-th station.

In testing for spatial alignment, we recommend using the following spatial correlation coefficients:

$$r_{x_j} = \frac{\frac{1}{K-1}\sum_{i=1}^{K}\left(P_{j,i} - \mu_{P_j}\right)\left(M_{j,i} - \mu_{M_j}\right)}{\sigma_{P_j}\sigma_{M_j}} \qquad \text{(VI-19)}$$

$$r_{x_{OVERALL}} = \frac{1}{N}\sum_{j=1}^{N} r_{x_j} \quad , \qquad \text{(VI-20)}$$

where r_{X_j} is the spatial correlation coefficient at the j-th hour for the K monitoring stations, and $r_{X_{OVERALL}}$ is the average correlation coefficient for all the N modeling period divisions (e.g., hours). Also, μ_{P_j} and σ_{P_j} are the mean and standard deviations of the predictions for K stations at the j-th hour. Similarly, μ_{M_j} and σ_{M_j} are the mean and standard deviations of the concentration at the j-th hour.

As for the form of the standard, we would require that

$$r \geq r_{min} \quad , \qquad \text{(VI-21)}$$

where r_{min} is defined at the 95 percent confidence level, perhaps using a t-statistic if no better method is apparent.

E. A SAMPLE CASE: THE SAI DENVER EXPERIENCE

In Section D we recommended a set of measures and standards for use in evaluating model performance. Here we illustrate how these measures might actually be used in practice. To do so, we draw on SAI experience in modeling the Denver metropolitan region (Anderson et al., 1977) using the grid-based SAI Airshed Model (Ames et al., 1978). We first show for the sample case the values we calculate for the performance measures; then we discuss how to interpret their meaning.

1. The Denver Modeling Problem

Over the past several years, Region VIII of the EPA has prepared an Overview EIS assessing the impact on the Denver metropolitan region of the proposed construction of twenty-two separate wastewater treatment projects. Adopting a regional approach, they assessed the projected impact of the facilities in several key ways, among which was their effect on air quality. They contracted with SAI in late 1976 to conduct that portion of the assessment. SAI employed several air quality models, one a long-term climatological model (CDM) and the other a short-term photochemical model (the SAI Airshed Model). We consider the latter of these in our sample case.

The grid-based Airshed Model is fully three-dimensional and capable of simulating concentrations of up to 13 chemical species, including ozone, nitrogen dioxide and several types of reactive hydrocarbons. The modeling grid chosen for overlaying the Denver Metropolitan region was 30 miles by 30 miles, subdivided horizontally into grid cells two miles on a side.

In cooperation with local agencies, SAI assembled meteorological information (spatial and temporal profiles of temperature and inversion height, as well as wind speeds and directions) characterizing atmospheric conditions on several summertime test days, 29 July 1975, 28 July 1976, and 3 August 1976. Also, gridded emissions inventories were compiled (hourly by species) for those days as were estimates for the years 1985 and 2000. Simulations were then conducted, with projections also made of air quality in the two subsequent years.

2. Values of the Performance Measures

We compare in this sample case the predicted and observed concentrations of ozone at each monitoring station in the regional measurement network. The issues we address are SIP/C and AQMP. On the test date we have chosen, 28 July 1976, eight monitoring stations provided ozone concentration data. Their locations are shown in Figure VI-2. Of the nine stations, all but CAMP provided usable ozone measurements. Data were recorded as hourly averages for each hour throughout the day.

The Airshed Model generates its predictions as grid cell-averaged hourly concentrations. Through interpolation, these values may then be used to estimate station predictions (concentrations at fixed points rather than grid cell averages). Plotted in Figure VI-3 are the predicted and observed ozone concentrations at each of the eight stations reporting on the modeled day (Anderson, et al., 1977). From the station predictions and observations, we can calculate performance measure values. We present the values of these measures in Table VI-12. We indicate in the table how these values might be interpreted in evaluating model performance, considering each in more detail below.

KEY

NG - Northglenn
WE - Welby
AR - Arvada
CR - C.A.R.I.H.
CM - Continuous Air Monitoring Program [CAMP]

NJ - National Jewish Hospital
GM - Green Mountain
OV - Overland
PR - Parker Road

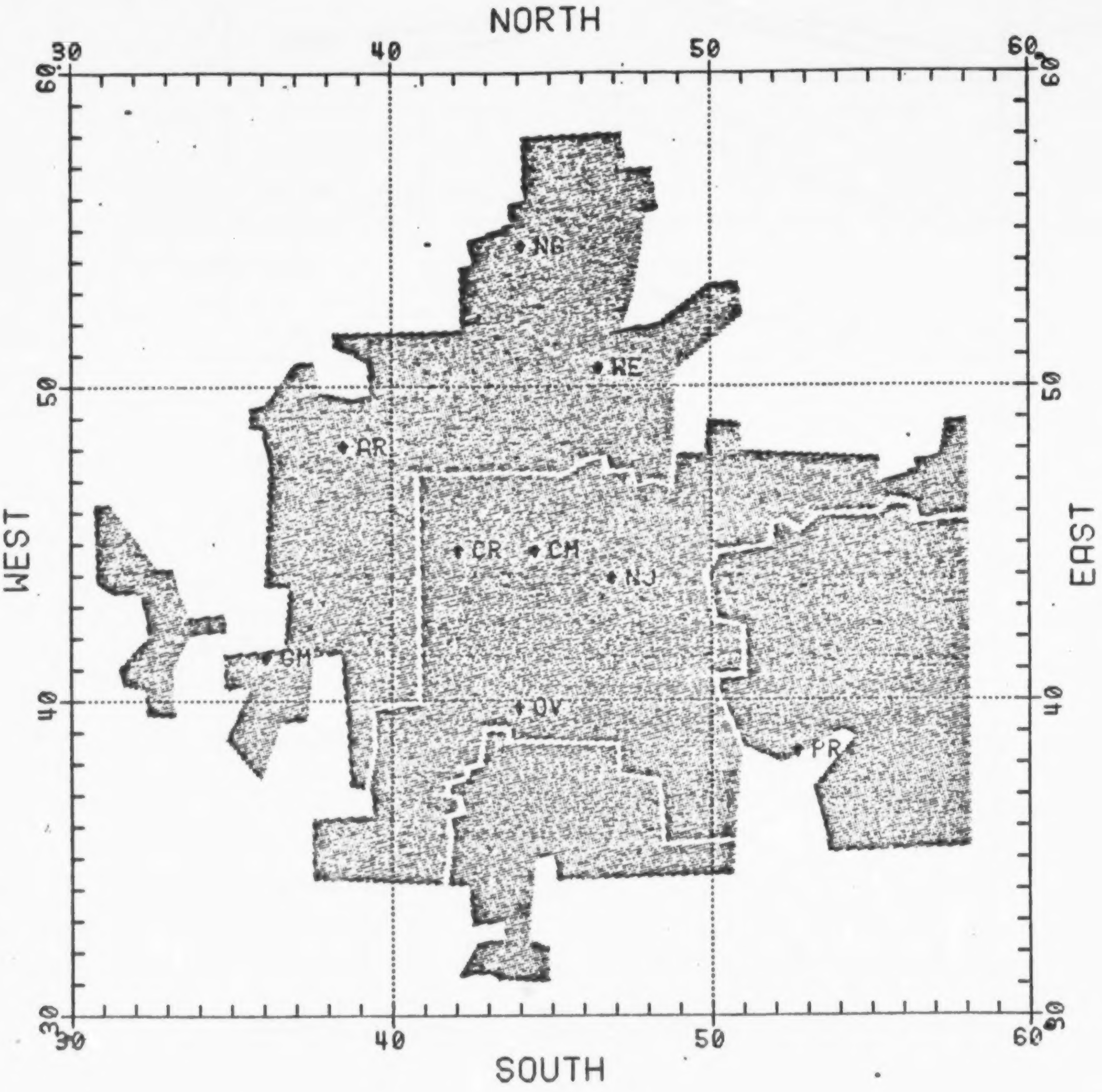

NORTH
SOUTH
WEST
EAST
30
40
50
60
NG
WE
AR
CR
CM
NJ
GM
OV
PR

FIGURE VI-2. LOCATIONS OF MONITORING STATIONS IN THE DENVER METROPOLITAN REGION

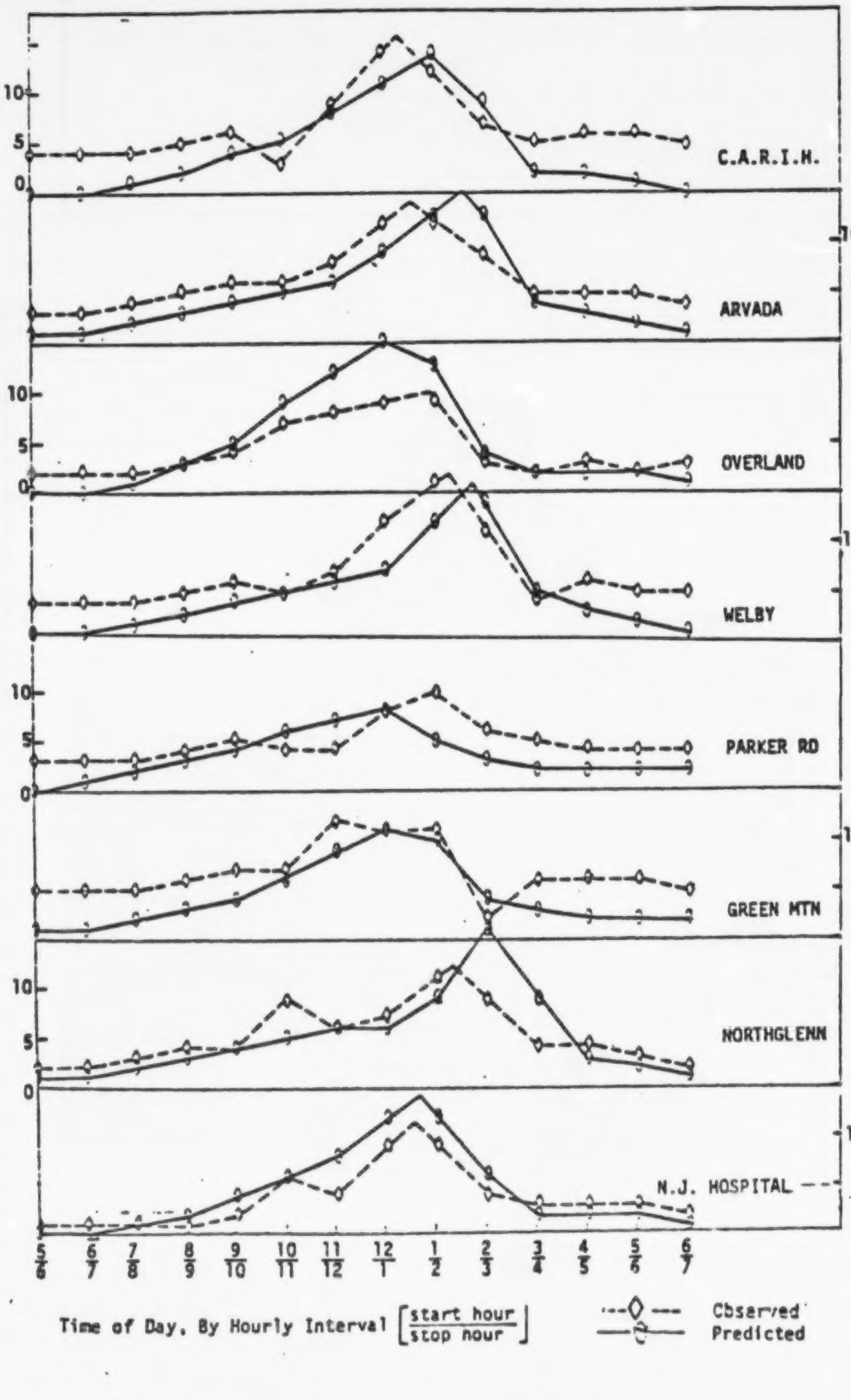


FIGURE VI-3. PREDICTED AND OBSERVED OZONE CONCENTRATIONS AT EACH MONITORING STATION DURING THE DAY (DENVER, 28 JULY 1976)

TABLE VI-12. SAMPLE VALUES FOR MODEL PERFORMANCE STANDARDS (DENVER EXAMPLE)

Performance Attribute	Composite Importance Category*	Performance Measure	Performance Standard	Calculated Value	Interpretation
Accuracy of the peak prediction	1	Ratio of predicted to measured station peaks C_{p_p}/C_{p_m}	$80 \leq \frac{C_{p_p}}{C_{p_m}} \leq 150$ percent	99 percent	Peak performance of the model is satisfactory.
		Timing of the peak[†] Δt_p	± 1 hour	+ 1 hour	The timing of the peak is satisfactory. Since the model provides only hourly averages, this is as finely as Δt_p can be determined.
Absence of systematic bias	1	Average value and standard deviation of the mean deviation about the perfect correlation line, normalized by the average of the predicted and observed concentrations $\bar{\mu}, \sigma_{\bar{\mu}}$	For concentrations (predicted or observed) at or above the NAAQS, the bias should not be greater than the maximum bias resulting from EPA-allowable monitor calibration error. A -8 percent bias--not normalized--is representative, which for this case is $\mu = -0.4$ pphm $\sigma_{\mu} = 1.53$ pphm for an EPA-acceptable monitor[§]--see Burton, et al. (1976)--when all concentrations are considered. An EPA-acceptable monitor can have an uncertainty with respect to a reference monitor of as much as ± 3 pphm for ozone at a 95 percent confidence level.	For concentrations greater than the NAAQS (8.0 pphm), $\bar{\mu} = 4.1\%$ $\sigma_{\bar{\mu}} = 19.4\%$. For all concentrations, $\bar{\mu} = -23.4\%$ $\sigma_{\bar{\mu}} = 33.5\%$. In a form suitable for comparison with non-normalized instrument bias, $\mu = -0.52$ pphm $\sigma_{\mu} = 1.22$ pphm when all concentrations are considered.	For concentrations at or above the NAAQS, a slight positive bias exists, though within acceptable bounds. When all concentrations are considered, a larger negative bias seems to exist. Put in a form suitable for comparison with an EPA-allowable monitor,[§] however, the bias appears to be indistinguishable from that resulting from maximum allowable calibration error. Overall, no conclusion of unacceptably high bias would seem justified.

TABLE VI-12 (Concluded)

Performance Attribute	Composite Importance Category*	Performance Measure	Performance Standard	Calculated Value	Interpretation
Lack of gross error	2	Average value and standard deviation of the absolute mean deviation about the perfect correlation line, normalized by the average of the predicted and observed concentrations $|\bar{u}|, \sigma_{|\bar{u}|}$	For concentrations at or above the NAAQS, the error should be indistinguishable from the distribution of error resulting from comparison of an EPA-acceptable monitor§ with a reference monitor. Representative values for an EPA-acceptable monitor (-8 percent bias; ± 3 pphm at a 95 percent confidence level) might be estimated to be $|u| = 1.22$ pphm $\sigma_{|u|} = 0.95$ pphm . Note that these values are based on non-normalized deviations.	For concentrations greater than the NAAQS (8.0 pphm), $|\bar{u}| = 16.7\%$ $\sigma_{|\bar{u}|} = 19.4\%$. For all concentrations, $|\bar{u}| = 31.5\%$ $\sigma_{|\bar{u}|} = 33.5\%$. In a form suitable for comparison with non-normalized instrument error, $|u| = 1.12$ pphm $\sigma_{|u|} = 0.72$ pphm .	For concentrations at or above the NAAQS, the error seems to be about half of what is seen if all concentrations are considered. The model thus appears to be subject to less error at the higher concentration range. We can determine the acceptability of this error level by converting to a non-normalized form for comparison with an estimate of that resulting from use of an EPA-acceptable monitor.§ Even when all concentrations are considered, the error in model predictions appears to be less than that resulting from monitoring technique differences. We conclude that the model performance is acceptably good insofar as error is concerned.
Temporal correlation	2	Temporal correlation coefficients at each monitoring station and an overall coefficient (the all-station average) $r_{t_i}, r_{t_{OVERALL}}$ for $1 \leq i \leq M$ monitoring stations	At a 95 percent confidence level, predicted and observed concentrations should appear to be correlated. Using a t-statistic to estimate the minimum acceptable correlation coefficient, in this example, we find $r_{t_{min}} = 0.53$.	For each monitoring station, $0.69 \leq r_{t_i} \leq 0.97$. Overall, $r_{t_{OVERALL}} = 0.88$.	For all stations and overall, predicted and observed concentrations appear to be correlated. The model performance appears to be within acceptable bounds.
Spatial alignment	2	Spatial correlation coefficients for each modeling hour and an overall coefficient for the entire day (the all-hours average) $r_{x_j}, r_{x_{OVERALL}}$ for $1 \leq j \leq N$ modeling hours	At a 95 percent confidence level, predicted and observed concentrations should appear to be correlated. Using a t-statistic to estimate the minimum acceptable correlation coefficient, in this example we find $r_{x_{min}} = 0.71$.	For each modeling hour, $-0.44 \leq r_{x_j} \leq 0.66$; . Overall, $r_{x_{OVERALL}} = 0.17$	During none of the hours considered (all daylight hours) do prediction and observation appear to be correlated at the 95 percent confidence level. Model predictions appear to be spatially misaligned, although the presence of temporal correlation suggests that the misalignment may not be a serious problem. (Another interpretation may be correct: Either r_x is too stringent a measure of spatial alignment or r_t is too lenient a measure of temporal behavior. Only by additional research, however, will we be able to confirm or refute this.)

---

* The composite importance category is determined by consulting Tables VI-2 and IV-3 for the appropriate issue and pollutant/averaging time (in this example, SIP/C and ozone/one-hour averaging time). The composite category is the less stringent of the two importance rankings.

† These measures are appropriate when the chosen model is used to consider questions involving photochemically reactive pollutants subject to short-term standards.

§ An "EPA-acceptable monitor" is defined here to be one that differs from a monitor using the EPA reference technique by up to the maximum allowable amount.

3. Interpreting the Performance Measure Values

Briefly, we summarize the conclusions suggested by the model performance measures. First, even though the predicted and observed concentration peaks occur at different monitoring stations and times (North Glenn at 2-3 p.m. versus Welby at 1-2 p.m.), their values agree quite closely, well within the acceptable tolerance.

Second, systematic bias appears to remain within acceptable limits. We can demonstrate this graphically, first by plotting prediction-observation pairs in a correlogram (see Figure VI-4) and then plotting the normalized mean deviation about the perfect correlation line as is done in Figure VI-5. From this latter figure (suggested by Anderson, et al., 1977) we see that the Airshed Model, while systematically underpredicting at concentration levels below 4.5 pphm, does not appear subject to such bias at concentrations above that level. Incidentally, recent internal studies at SAI have indicated that the Denver region may be subject to background concentrations as high as 4 pphm (Anderson, 1978), values substantially higher than those supplied as input to the Airshed Model. Also, we may compare the deviations about the perfect correlation line to those that we would expect from comparison of an EPA-acceptable monitor with a monitor using the EPA reference technique (normally distributed, -8 percent bias, ± 3 pphm at the 95 percent confidence level--see Burton, et al., 1976). This comparison is shown in Figure VI-6. To aid in presenting this graphical comparison, we have converted deviations to the non-normalized form. We observed that the means (a measure of systematic bias) of both are nearly the same and that the standard deviation of prediction-observation deviations is somewhat less than that of the monitoring error distribution.

Third, consistent with our conclusions about systematic bias, gross error also appears to be within tolerable bounds. We show in Figure VI-7 the distribution of non-normalized error, that is, the absolute deviation of predictions and observations from the perfect correlation line. For reference we also estimate the corresponding distribution resulting from

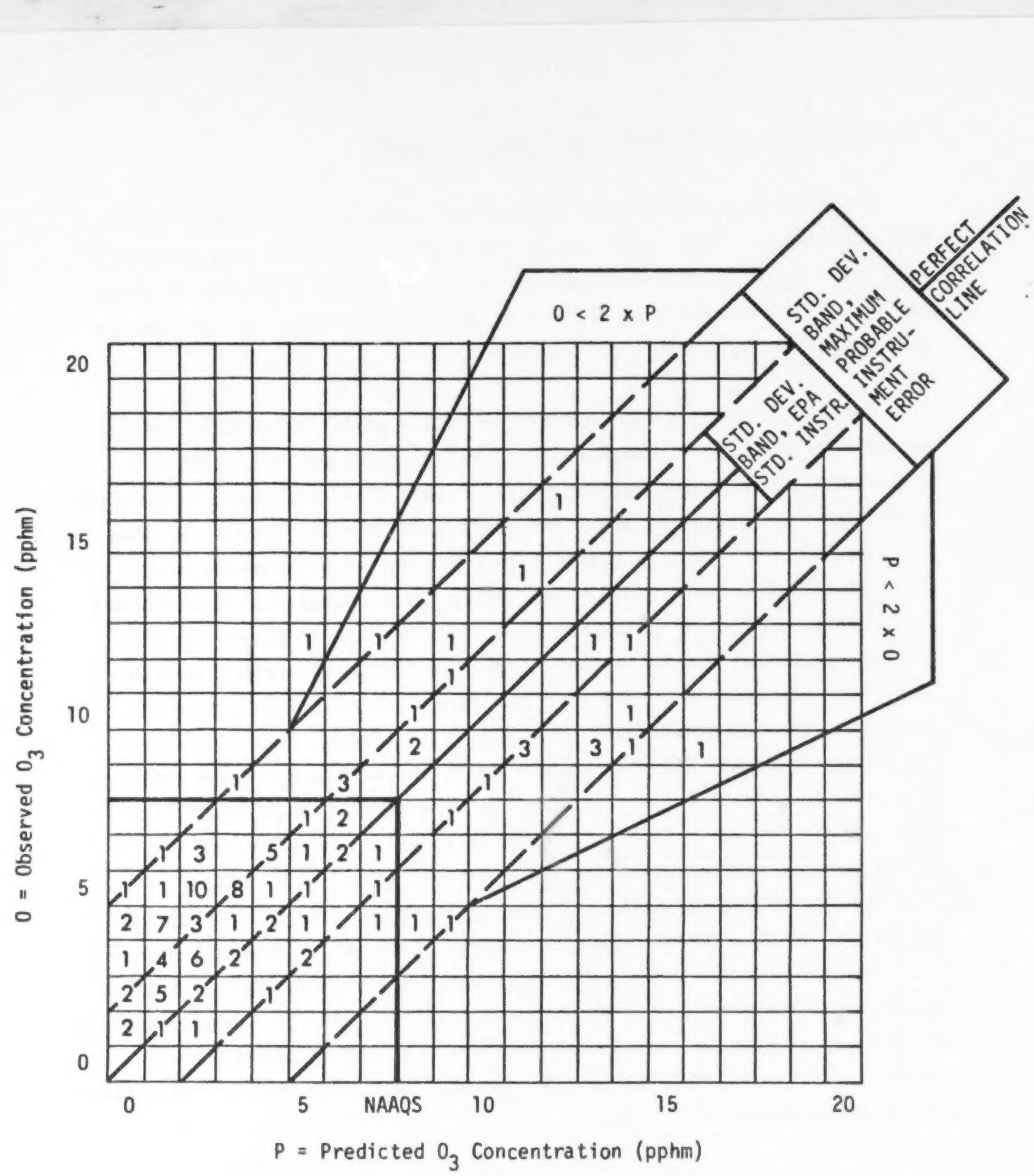


FIGURE VI-4. CORRELOGRAM OF OZONE OBSERVATION-PREDICTION PAIRS FOR SAMPLE CASE (DENVER, 28 JULY 1976)

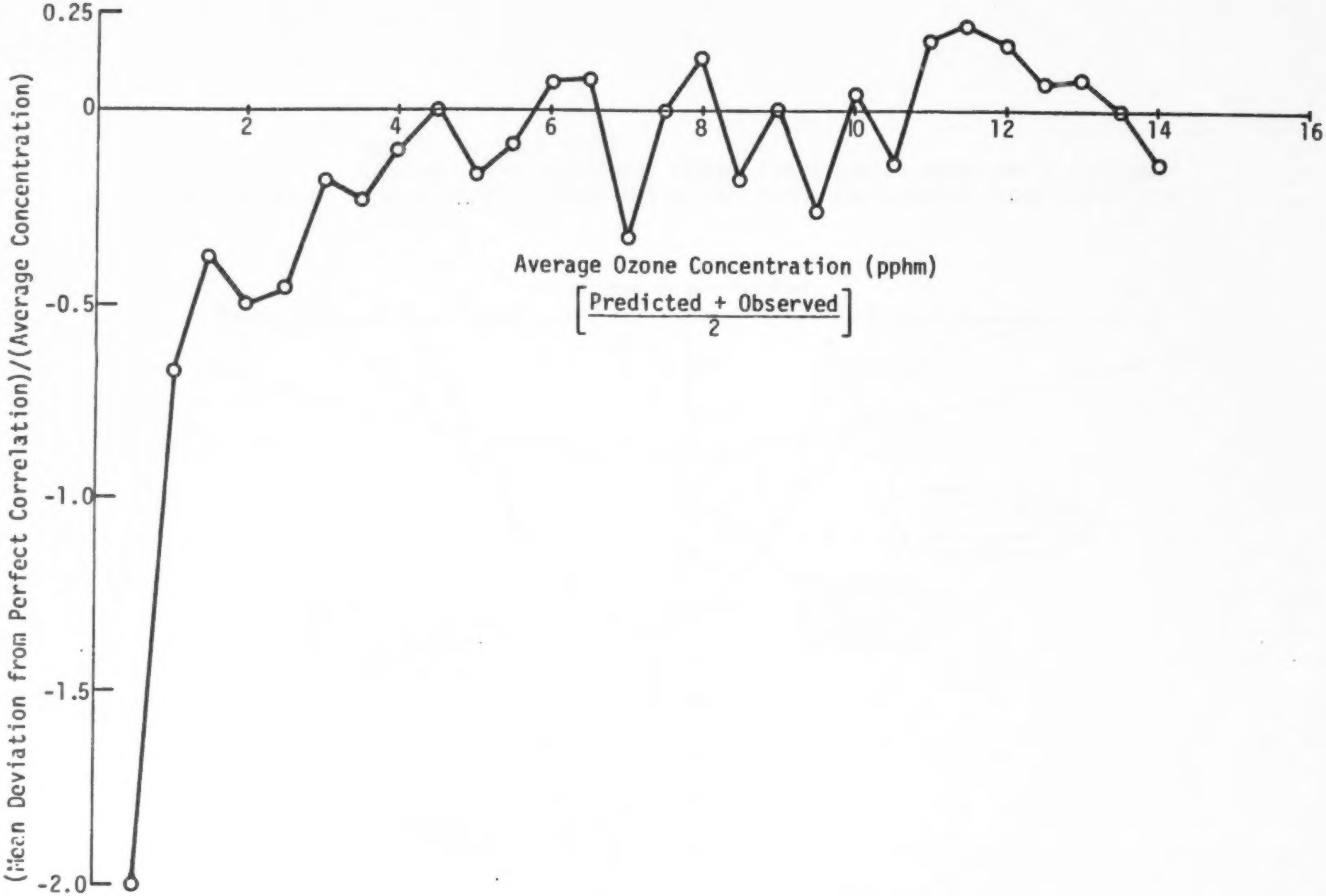


FIGURE VI-5. NORMALIZED DEVIATIONS ABOUT THE PERFECT CORRELATION LINE AS A FUNCTION OF OZONE CONCENTRATION (DENVER, 28 JULY 1976)

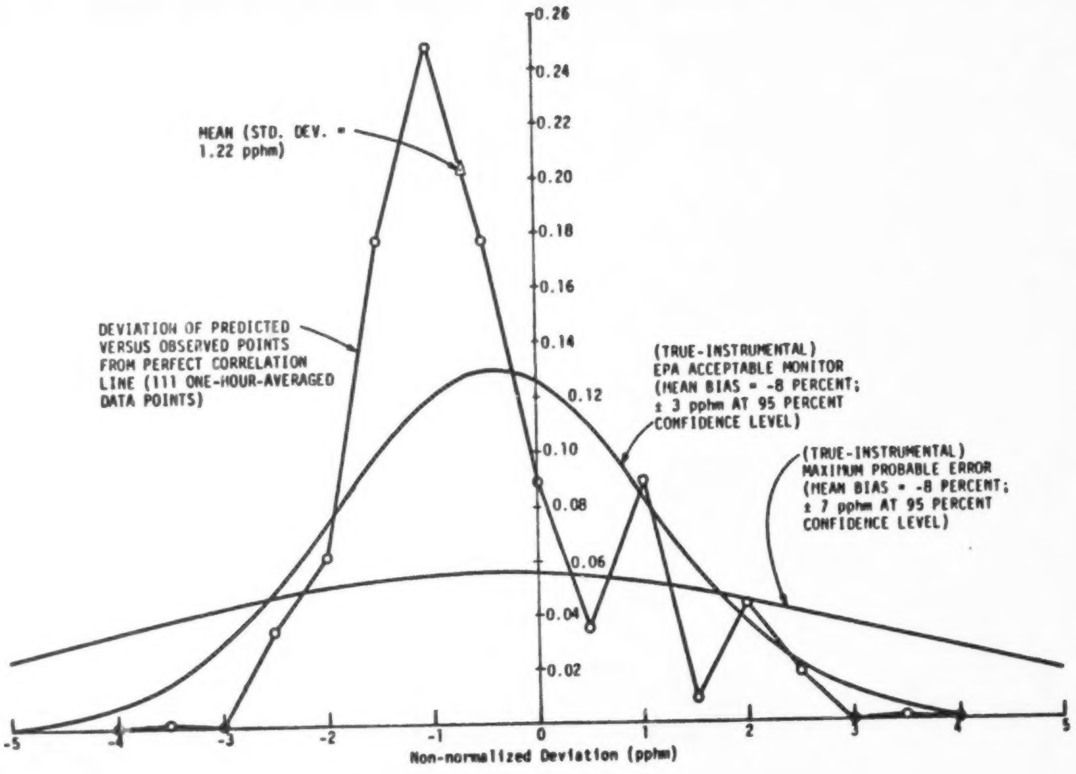


FIGURE VI-6. NON-NORMALIZED OZONE DEVIATIONS ABOUT THE PERFECT CORRELATION LINE COMPARED WITH INSTRUMENT ERRORS (DATA FOR 14 HOURS AND 8 STATIONS, DENVER, 28 JULY 1976)

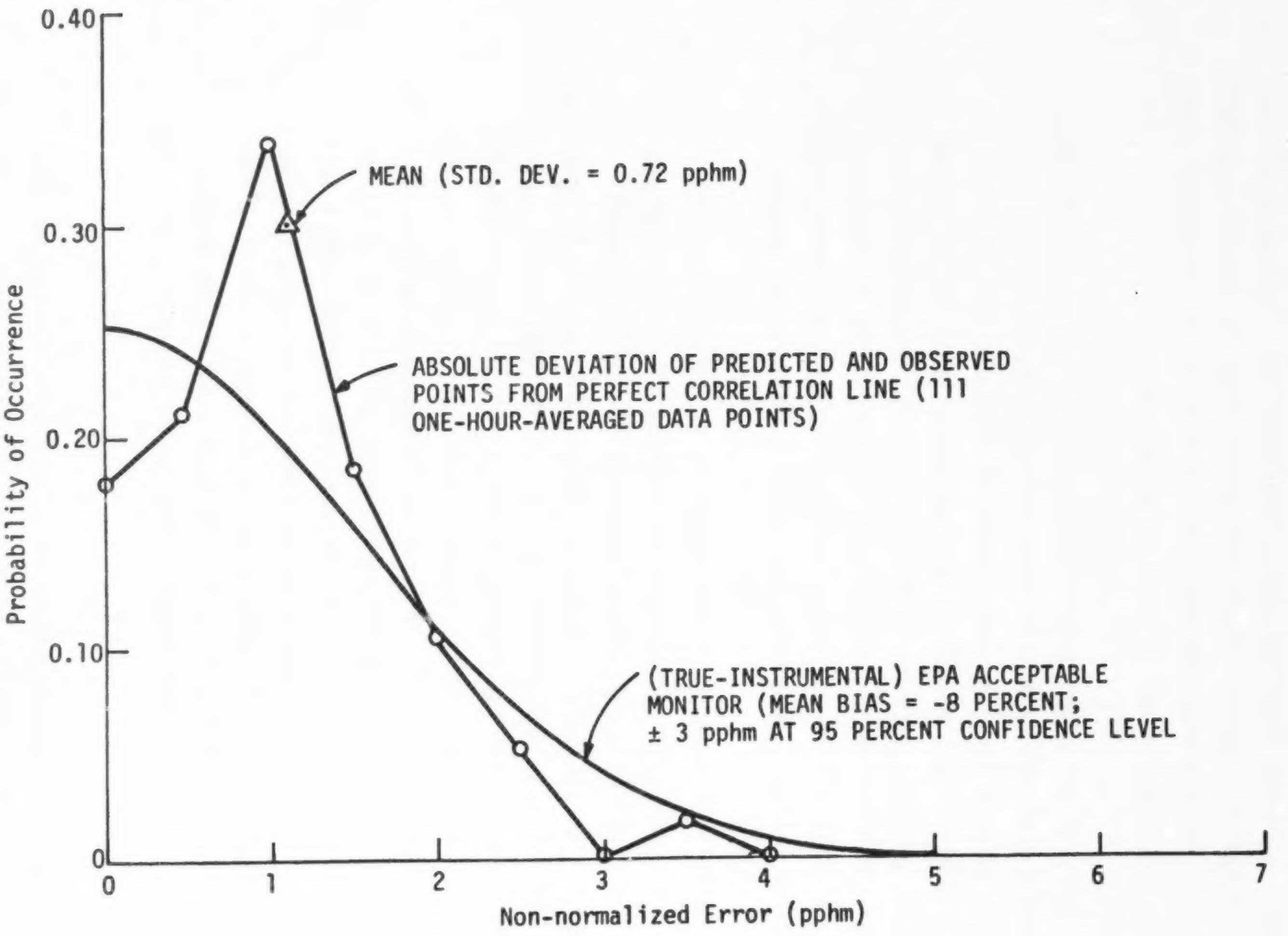


FIGURE VI-7. NON-NORMALIZED OZONE ABSOLUTE DEVIATIONS ABOUT THE PERFECT CORRELATION LINE COMPARED WITH INSTRUMENT ERROR (DATA FOR 14 HOURS AND 8 STATIONS, DENVER, 28 JULY 1976)

comparison of an EPA-acceptable monitor with an EPA reference instrument. We see that the mean value and standard deviation of the prediction-observation "error" are both somewhat less than those resulting from instrument differences. The conclusion suggests itself that gross error is within acceptable bounds, though we caution that the shape of the instrument difference curve is an estimate and needs to be analyzed in further detail.

Fourth, temporal behavior at each monitoring station seems satisfactory, appearing correlated to better than the requisite 95 percent confidence level. We note that the correlation we have observed provides information only about the "shape" of the concentration profiles (shown in Figure VI-3), not its absolute level. In general, predicted concentrations rise and fall when observed values do, though the concentration values might be quite different. Only by examining bias and error performance measures can we draw conclusions about concentration levels.

Fifth, spatial alignment does not appear to be acceptably good. During none of the 14 hours considered, do the spatial patterns of predictions and observations appear to be correlated at the 95 percent confidence level. In fact, for a number of hours, the correlation seems quite poor. Two possible explanations exist. Either the spatial correlation coefficient is too "stringent" or the predicted concentration field in fact is misaligned. Since temporal correlation appears strong, the lack of corresponding spatial correlation is somewhat surprising, though countervailing errors responsible for this conceivably could be present. It is also possible that the temporal correlation coefficient either is too "lenient" or it should not be computed including concentrations at all daylight hours. Presently, we do not know which of these explanations is correct, noting only that it is a subject for future investigation. Conceivably, measurement data errors could also be contributing to the problem.

In this example, we can examine model predictions for spatial misalignment. To do so, we conducted an informal experiment among several of our staff. In general, reconstructing the "true" concentration field from a "sparse" set of observational data is a difficult and uncertain process. Nevertheless, we attempted, using only station measurement data, to draw isopleth maps showing contours of constant concentration values. The process, of course, is a highly subjective one, requiring the person doing the drawing to make a number of judgmental and often arbitrary decisions. In this case, a useful result was achieved.

None of the participants in the experiment were able to draw unambiguous isopleth maps for those hours when overall concentrations were low (before 11 in the morning and after 3 in the afternoon). However, while they varied widely in their estimates during the four "peak hours" of the configurations for lower outlying concentration isopleths, each agreed reasonably well on their estimates of the location of the peak. We compare in Figure VI-8 a "ground-trace" of their composite estimates with the peak locations predicted by the Airshed Model.

We observe that the ground-traces of the predicted and observed peaks differ, both in direction and speed of drift. This suggests that either the model has had some difficulty in simulating atmospheric dispersion or it is being driven by inputs that imperfectly characterize ambient conditions on the modeling day. Based on a generally favorable model performance rating, as judged by the other four types of measures, we feel the latter of these two explanations is more likely.

The model input data most likely to have caused the alignment problem is the temporally and spatially varying wind field. By comparing the ground-trace of the predicted peak with the directions and speeds of prevailing winds that we input to the Airshed model, we confirmed that the wind field did indeed appear to be "forcing" the predicted pollutant cloud in just the direction noted in Figure VI-8.

We emphasize that this does not confirm that "errors" in the input wind field were responsible for the spatial misalignment, but the evidence is suggestive. Final confirmation or refutation would come by

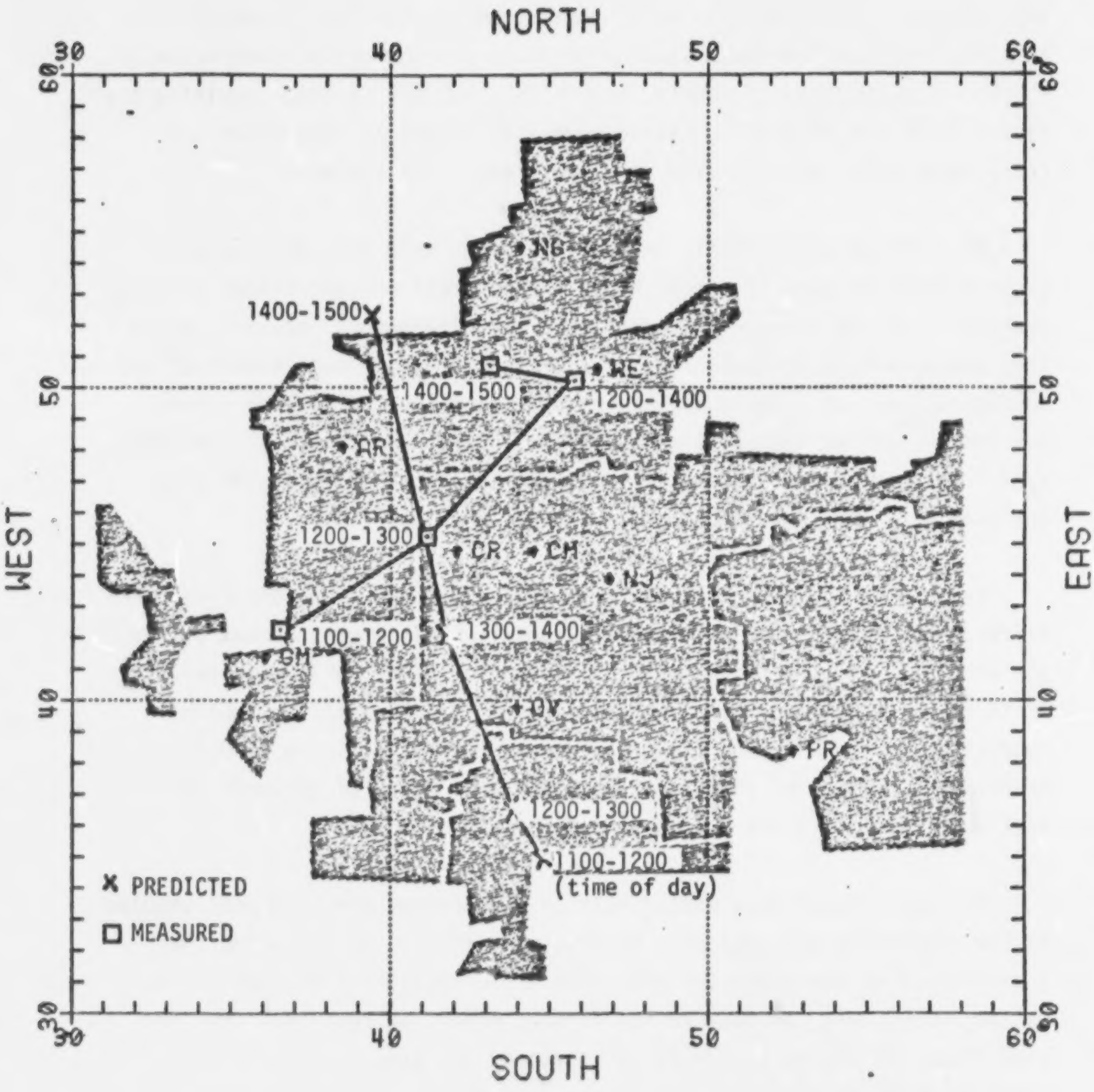

NORTH
SOUTH
WEST
EAST
30
40
50
60
NB
1400-1500
1400-1500
WE
1200-1400
AR
1200-1300
CR
CM
NJ
1100-1200
1300-1400
GM
DV
FR
1200-1300
1100-1200
(time of day)
✕ PREDICTED
☐ MEASURED

FIGURE VI-8. GROUND-TRACES OF THE PREDICTED AND OBSERVED PEAK OZONE CONCENTRATIONS (DENVER, HOURS 1100-1200 TO 1400-1500 LOCAL STANDARD TIME, 28 JULY 1976)

rerunning the Airshed Model using a wind field "adjusted" to better mirror our updated estimates of the meteorology on the modeling day. If agreement, as evaluated by the five types of performance measures, were "better," then we might conclude that wind field imperfections were responsible for our misalignment problems.

F. SUGGESTED FRAMEWORK FOR A DRAFT STANDARD

We have now completed our central objective in this report: the identification and specification of model performance measures and standards. In doing so, however, we have not solved the problem but rather only begun a discussion that will be a continually evolving one. Almost certainly, the specific measures and standards employed to evaluate model performance will change as our insight and experience expands. On balance, the most enduring benefit from this study will be the conceptual structure it sets.

With that structure in mind, we discuss one final subject: a framework for a draft model performance standard. We view the promulgation of the standard as having two distinct parts: the text of the standard itself and an accompanying guidelines document. Where the standard should be quite specific about selecting and applying the performance measures to be used, there needs to be a guidelines document in which supplementary discussion and examples are provided. While a full examination of the interrelationships between the two documents is beyond the scope of the current study, we illustrate in Figure VI-9 one possible configuration.

We focus in this discussion on suggested elements of a draft performance standard. We state several of the functional sections it should contain:

> Goals and Objectives. The reasons for insisting on model validation should be stated, as well as a summary of expected costs and benefits. Our objectives in conducting performance evaluation should be clearly presented.

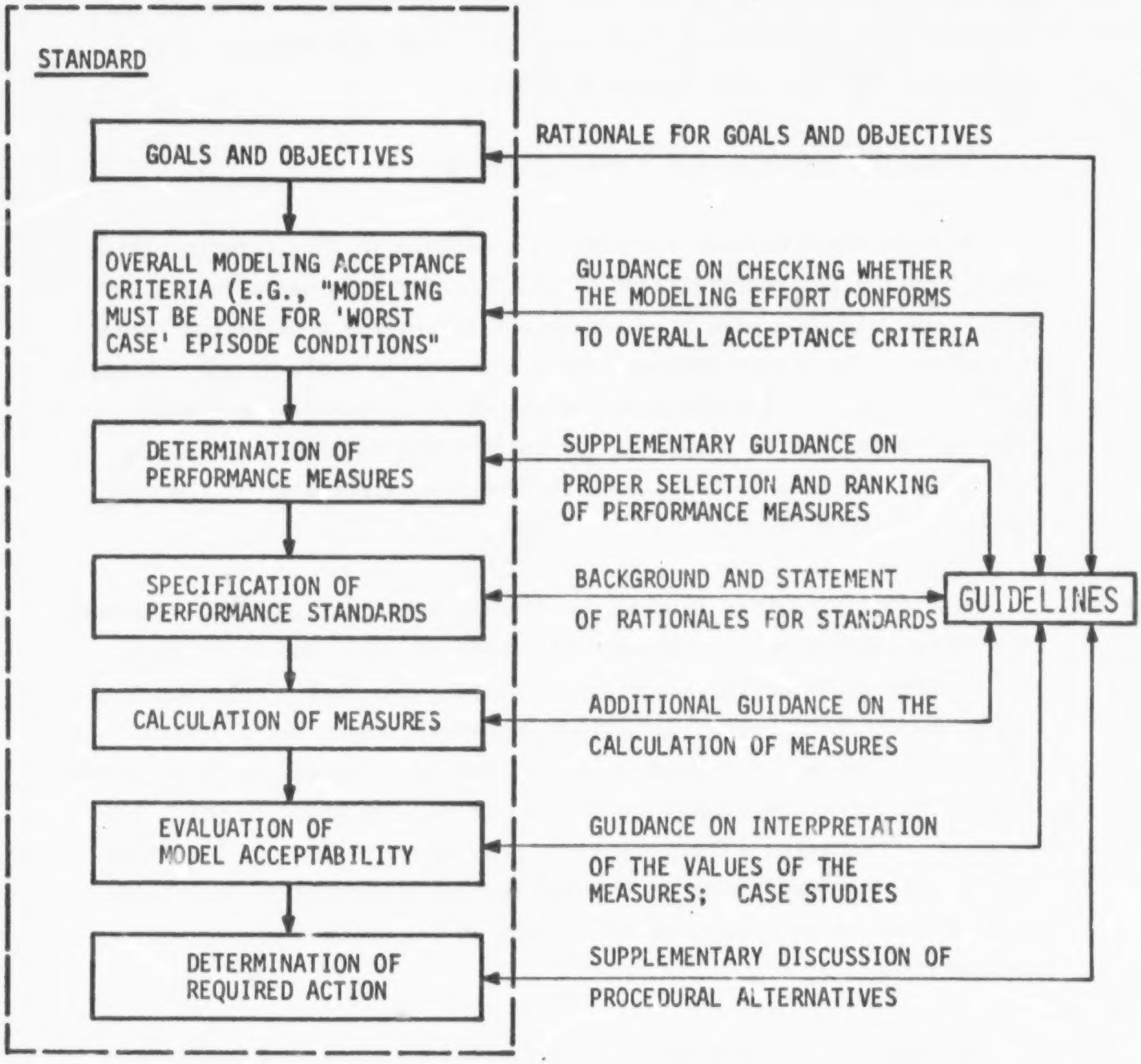

STANDARD
GOALS AND OBJECTIVES
OVERALL MODELING ACCEPTANCE CRITERIA (E.G., "MODELING MUST BE DONE FOR 'WORST CASE' EPISODE CONDITIONS"
DETERMINATION OF PERFORMANCE MEASURES
SPECIFICATION OF PERFORMANCE STANDARDS
CALCULATION OF MEASURES
EVALUATION OF MODEL ACCEPTABILITY
DETERMINATION OF REQUIRED ACTION
RATIONALE FOR GOALS AND OBJECTIVES
GUIDANCE ON CHECKING WHETHER THE MODELING EFFORT CONFORMS TO OVERALL ACCEPTANCE CRITERIA
SUPPLEMENTARY GUIDANCE ON PROPER SELECTION AND RANKING OF PERFORMANCE MEASURES
BACKGROUND AND STATEMENT OF RATIONALES FOR STANDARDS
ADDITIONAL GUIDANCE ON THE CALCULATION OF MEASURES
GUIDANCE ON INTERPRETATION OF THE VALUES OF THE MEASURES; CASE STUDIES
SUPPLEMENTARY DISCUSSION OF PROCEDURAL ALTERNATIVES
GUIDELINES

FIGURE VI-9. POSSIBLE RELATIONSHIPS BETWEEN THE MODEL PERFORMANCE STANDARDS AND A GUIDELINES DOCUMENT

> Overall Modeling Acceptance Criteria. Important criteria for judging a modeling effort in an overall sense should be clearly stated, along with the action required if any of the criteria are not satisfied. Among possible criteria are the following: The verification must be done for modeling days typical of "worst case" conditions, the measurement network must meet certain stated minimum standards (numbers, types and configurations of the monitoring stations), and point source models must be verified using the appropriate prototypical data base (one appropriate for an application similar to the proposed hypothetical one). Without these and perhaps other overall criteria being satisfied, model evaluation would be premature.

> Determination of Performance Measures. The procedure must be stated for determining the performance measures to be used for model evaluation. Instructions must also be provided for matching the importance ranking of each of the model performance attributes to the type of issue being addressed and the pollutant/averaging time being considered. We might do so using the importance tables we presented earlier in this chapter and repeat for convenience in Tables VI-13 and VI-14.

> Specification of Performance Standards. The standards must be clearly stated for each of the performance measures to be used. We present in Table VI-15 one format for doing so, presenting the standards in the form of general principles. In each instance, the actual numerical standard is dependent on the characteristics of the specific application. Guidance must be provided on how to determine the proper numerical values.

> Calculation of Measures. Each measure should be defined mathematically, accompanied by directions on precisely how the measures are to be calculated.

TABLE VI-13. IMPORTANCE OF PERFORMANCE ATTRIBUTES BY ISSUE

	Importance of Performance Attribute*						
Performance Attribute	**SIP/C**	**AQMP**	**PSD**	**NSR**	**OSR**	**EIS/R**	**LIT**
Accuracy of the peak prediction	1	1	1	1	2	1	1
Absence of systematic bias	1	1	1	1	1	1	1
Lack of gross error	2	2	1	1	2	1	1
Temporal correlation	2	2	3	3	3	3	3
Spatial alignment	2	2	1	3	3	3	3

---

* Category 1 - Performance standard must always be satisfied.
Category 2 - Performance standard should be satisfied, but some leeway may be allowed at the discretion of a reviewer.
Category 3 - Meeting the performance standard is desirable but failure is not sufficient to reject the model; measures dealing with this problem should be regarded as "informational."

TABLE VI-14. IMPORTANCE OF PERFORMANCE ATTRIBUTES BY POLLUTANT AND AVERAGING TIME

	Importance of Performance Attribute*										
	Pollutants with Short-term Standards								Pollutants with Long-term Standards		
Performance Attribute	O_3^{**} (1 hour)§	CO** (1 hour)	NMHC* (3 hour)	SO_2 (3 hour)	NO_2 (?)†	CO (8 hour)	TSP** (24 hour)	SO_2^{**} (24 hour)	NO_2^{**} (1 year)	TSP (1 year)	SO_2 (1 year)
Accuracy of the peak prediction	1	1	1	1	1	1	1	1	3	3	3
Absence of systematic bias	1	1	1	1	1	1	1	1	1	1	1
Lack of gross error	1	1	1	1	1	1	1	1	1	1	1
Temporal correlation	1	2	2	2	1	2	3	3	N/A††	N/A	N/A
Spatial alignment	1	2	2	2	1	2	2	2	2	2	2

---

* Category 1 - Performance standard must be satisfied.
Category 2 - Performance standard should be satisfied, but some leeway may be allowed at the discretion of a reviewer.
Category 3 - Meeting the performance standard is desirable but failure is not sufficient to reject the model.

† No short-term NO_2 standard currently exists.

§ Averaging times required by the NAAQS are in parentheses.

** Primary standards.

†† The performance attribute is not applicable.

TABLE VI-15. MODEL PERFORMANCE MEASURES AND STANDARDS[*]

Performance Attribute	Performance Measure	Performance Standard
Accuracy of the peak prediction	Ratio of the predicted station peak to the measured station (could be at different stations) C_{p_p}/C_{p_m}	Limitation on uncertainty in aggregate health impact and pollution abatement costs[*]
	Difference in timing of occurrence of station peak[†] Δt_p	Model must reproduce reasonably well the phasing of the peak--say, ±1 hour
Absence of systematic bias[§]	Average value and standard deviation of the mean deviation about the perfect correlation line, normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either the predicted or the observed values exceed some appropriate minimum value (possibly the NAAQS) $(\bar{\mu}, \sigma_{\bar{\mu}})_{\text{OVERALL}}$	No or very little systematic bias at concentrations (predictions or observations) at or above some appropriate minimum value (possibly the NAAQS); the bias should not be worse than the maximum bias resulting from EPA-allowable calibration error (-8 percent is a representative value for ozone); also, the standard deviation should be less than or equal to that of the difference distribution between an EPA-acceptable monitor[**] and an EPA reference monitor (3 pphm is representative for ozone at the 95 percent confidence level)
Lack of gross error[§]	Average value and standard deviation of the absolute mean deviation about the perfect correlation line, normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either the predicted or the observed values exceed some appropirate minimum value (possibly the NAAQS) $(\lvert\bar{\mu}\rvert, \sigma_{\lvert\bar{\mu}\rvert})_{\text{OVERALL}}$	For concentrations at or above some appropriate minimum value (possibly the NAAQS) the error (as measured by the overall values of $\lvert\bar{\mu}\rvert$ and $\sigma\lvert\bar{\mu}\rvert$) should not be worse than the error resulting from the use of an EPA-acceptable monitor[**]
Temporal Correlation[†]	Temporal correlation coefficients at each monitoring station for the entire modeling period and an overall coefficient averaged for all stations $r_{t_i}, r_{t_{\text{OVERALL}}}$, for $1 \leq i \leq M$ monitoring stations	At a 95 percent confidence level, the temporal profile of predicted and observed concentrations should appear to be in phase (in the absence of better information, a confidence interval may be converted into a minimum allowable correlation coefficient by using an appropriate t-statistic)
Spatial alignment[†]	Spatial correlation coefficients calculated for each modeling hour considering all monitoring stations, as well as an overall coefficient average for the entire day $r_{x_j}, r_{x_{\text{OVERALL}}}$, for $1 \leq j \leq N$ modeling hours	At a 95 percent confidence level, the spatial distribution of predicted and observed concentrations should appear to be correlated

---

[*] There is deliberate redundancy in the performance measures. For example, in testing for systematic bias, $\bar{\mu}$ and $\sigma_{\bar{\mu}}$ are calculated. The latter quantity is a measure of "scatter" about the perfect correlation line. This is also an indicator of gross error and should be used in conjunction with $\lvert\bar{\mu}\rvert$ and $\sigma_{\lvert\bar{\mu}\rvert}$.

[§] These measures are appropriate when the chosen model is used to consider questions involving photochemically reactive pollutants subject to short-term standards.

[†] These may not be appropriate for all regulated pollutants in all applications. When they are not, standards derived based on pragmatic/historic experience should be employed.

[**] By "EPA-acceptable monitor" we mean a monitor that satisfies the requirements of 40 CFR §53.20.

- Evaluation of Model Acceptability. The rating procedure to be used in evaluating model performance must be stated. Guidance should be supplied on the way in which problem importance ranking is "folded in" with the performance rating for each of the measures.
- Determination of Required Action. The alternative actions required of the model user, depending on the model evaluation, must be stated. Among the possible alternative outcomes of the model evaluation are the following: The model is rated acceptable, the model requires a waiver from an outside reviewer before acceptance can be granted (that is, the model is deficient in some Category 2-importance problem area), or the model is unacceptable (the model is deficient in some Category 1-importance problem area).

We end our discussion of a suitable structure for a draft performance standard by noting that this has been only a brief encounter with an important and complex subject. We recommend that it be examined in far greater detail in subsequent work.

VII RECOMMENDATIONS FOR FUTURE WORK

In this study we have suggested a conceptual framework within which model performance may be objectively evaluated. We have identified key attributes of a well performing model and selected performance measures for use in detecting the presence or absence of each attribute. For the measures chosen for use, we have developed explicit standards that specify the range of their acceptable values.

Throughout, we have maintained the point of view that measures and standards of performance for models should be determined as independently as possible of considerations about model-specific limitations and data inadequacies. Remembering this perspective may be important when evaluating the practical utility of the procedure suggested in this report in certain point source applications. This is particularly true when the available measurement data are "sparse." Where data quantity and resolution (temporal and spatial) are insufficient to permit meaningful calculation of the performance measures, we view this more as a data inadequacy that must be overcome than as a deficiency in the model evaluation framework suggested here.

The development of a performance evaluation procedure for models is an evolutionary process. We have advanced in this study a conceptual structure and a first-generation procedure for conducting such an evaluation. We now recommend ways in which development may proceed, moving from the conceptual framework provided in this study to the realm of practical application of performance evaluation procedures.

We recommend that the work begun in this study continue in several key areas. In this chapter we outline briefly our specific recommendations, grouping them into three categories: areas for technical development, assessment of institutional implications, and documents to be compiled. We consider each category in turn.

A. AREAS FOR TECHNICAL DEVELOPMENT

A number of important technical areas remain that would benefit from additional developmental work. We consider four key areas here.

1. Further Evaluation of Performance Measures

In this study, a sample case has been considered that permits us to evaluate in a practical situation the utility of the recommended performance measures in detecting the presence or absence of desirable model attributes. However, the suitability for use of each of these measures needs further evaluation over a range of circumstances. Specifically, we recommend the following:

- > Additional case studies need to be considered, with performance measures calculated for each. The choice of case studies should be made in order to "stress" the evaluation procedure, that is, any limitations should be made apparent. The range of case studies should include both multiple-source and specific-source applications.
- > The behavior of the suggested performance measures needs to be assessed over a range of conditions. Alternate or supplementary performance measures should be identified, if required, so as to further extend the range of applicability of the evaluation procedure suggested in this study.
- > A performance measure evaluation analysis should be conducted. Two concentration fields, initially aligned spatially and temporally, could be progressively "degraded," that is, offset in space or time. By observing the corresponding changes in the values of the performance measures and the conclusions that derive therefrom, insight could be gained into their overall suitability for use.

2. Identification and Specification of Prototypical Point Source "Test Bed" Data Bases

For the purposes of model evaluation in the many specific-source applications where site-specific data are either inadequate or nonexistent, a "test-bed," or surrogate, data base is required. This data base must provide concentration data of sufficient spatial extent and temporal frequency to permit the calculation of meaningful values for the model performance measures. Selection of a particular data base could be made by determining, from among several prototypical "test beds," which derives from conditions most like those in the proposed application. We recommend that the following work be undertaken:

- A comprehensive list of prototypical point source situations should be compiled.
- For each prototypical situation, a "test bed" data base should be specified and assembled.

3. Examination of Performance Evaluation Procedure in Sparse-Data Point Source Applications

We have identified in this study several key attributes of a well-performing model, for each of which presence or absence may be detected by calculating certain performance measures. However, for the values of these measures to assume statistical significance, a certain minimum level is required for the spatial extent and the temporal frequency of the measurement data. Often, in multiple-source applications, such a minimum level is attained, particularly in urban areas with well-developed monitoring networks. In specific-source applications, though, a minimum acceptable level of data may not be attained. To overcome this problem, we have suggested that prototypical point source data bases be assembled for the purposes of model evaluation. These data bases would provide sufficiently well-conditioned data for calculation of the performance measures to be useful.

As a practical matter, however, such data bases are not presently available to the modeling community. In lieu of their use, other sources of data may be used for the purpose of model evaluation, despite the deficiencies in such data. For example, a limited amount of tracer data may be gathered. If the situation to be modeled involves either construction at a site where another source already exists or retrofit of pollution control equipment, then some limited site-specific monitoring data may be available. Such data may not be sufficiently "well-conditioned" to permit meaningful calculation of the performance measures suggested for use. What can be done? Should calculation of the performance measures be allowed using the possibly deficient, sparse data available, or should the model evaluation process be halted until more "robust" data are acquired? We suggest that the implications of both these alternatives be assessed, searching for those limited circumstances where a "middle ground" may be found, with alternative measures and standards identified for use that are less "demanding" in their measurement data requirements. The implications of allowing the use of such supplementary measurements also need to be examined.

Also, a related issue may be important in point source modeling applications: relative versus absolute model performance. Are there circumstances in which a model may be better able to predict relative, incremental changes in concentration than absolute ground-level values? It should be determined whether or not such situations occur in practice. If they do, relative validation of a model may become a consideration. This could be of concern, for example, when using a Gaussian model to assess the impact of control equipment that is retrofit on an existing source. If relative performance is deemed important in some circumstances, then additional performance measures and standards should be identified which allow the modeler to make such an assessment.

4. Further Development of Rationales for Setting Performance Standards

Several rationales for setting performance standards have been examined in this study. Some of these merit further technical development and assessment of the range of their applicability. Also, additional rationales should be identified where possible. Towards these ends, we recommend the following:

> Additional developmental work should continue on the Health Effects (HE) and Control Level Uncertainty (CLU) rationales.

> The use of the HE/CLU rationales in setting a standard for the ratio of predicted and observed peak station concentra-

tions should be exposed to peer review. A journal article on the subject should be prepared and submitted for publication.

- Explicit error and bias standards should be calculated for all regulated pollutants. This may be done using monitoring specifications in federal regulations. In this study, only bias and error standards for ozone were calculated numerically.

B. ASSESSMENT OF INSTITUTIONAL IMPLICATIONS

A number of institutional requirements are implied by any decision to promulgate standards for model performance, or even by a decision to publish formal guidelines for model performance evaluation. We recommend that these implications and their attendant procedural and resource requirements be assessed. Among the many questions to be resolved are the following:

- Regulatory Responsibility
  - How should formal performance standards be promulgated--or should they be promulgated at all?
  - If standards are stated or recommended, how will they be updated?
  - Who will accumulate information about historically achieved model performance? (This information would be required when setting a standard invoking the Pragmatic/Historic rationale.)
- Custodial Responsibility
  - Who will identify and assemble the prototypical "test bed" data bases for use in point source applications?
  - Who will maintain, store, and distribute the "test bed" data bases?
- Review Responsibility
  - Who should review the adequacy of model performance in a specific application?

  - Does a model need to be repeatedly evaluated using a "test bed" data base? If not, who decides when a model/data base combination has been sufficiently examined?
- > Advisory Responsibility
  - What advisory documents should be provided to the model user community?
  - Who will provide guidance to model users and how should that support be funded?

These are simply a few of the many procedural and institutional questions that arise. Answers to these and other key questions should be sought at an early date.

C. DOCUMENTS TO BE COMPILED

Specific documents will have to be drafted that describe suggested or mandated model performance standards. Two documents seem appropriate for publication (though conceivably they could be combined into a single guidelines document). These documents are the following:

- > Formally promulgated model performance standards along with specific procedures for evaluating performance. These could be presented in guideline form rather than as mandated standards. The latter of these two approaches may be preferable, given the complexities of modeling and its attendant uncertainties.
- > Advisory/informative model performance guidelines document. This may provide the advice and information necessary to conduct a meaningful model performance evaluation. It could play the role, with respect to the performance standards, that is indicated in Figure VI-9.

APPENDIX A

IMPORTANT PARTS OF THE CODE OF FEDERAL REGULATIONS CONCERNING AIR PROGRAMS

APPENDIX A

IMPORTANT PARTS OF THE CODE OF FEDERAL REGULATIONS CONCERNING AIR PROGRAMS

PART 50. NATIONAL PRIMARY AND SECONDARY AMBIENT AIR QUALITY STANDARDS

Section	
50.1	Definitions.
50.2	Scope.
50.3	Reference Conditions.
50.4	National primary ambient air quality standards for sulfur oxides (sulfur dioxide).
50.5	National secondary ambient air quality standards for sulfur oxides (sulfur dioxide).
50.6	National primary AAQS for particulate matter.
50.7	National secondary AAQS for particulate matter.
50.8	National primary and secondary AAQS for carbon monoxide.
50.9	National primary and secondary AAQS for photochemical oxidants.
50.10	National primary and secondary AAQS for hydrocarbons.
50.11	National primary and secondary AAQS for nitrogen dioxide.

Appendix A--Reference Method for the Determination of Sulfur Dioxide in the Atmosphere (Pararosaniline Method).

Appendix B--Reference Method for the Determination of Suspended Particulates in the Atmosphere (High Volume Method).

Appendix C--Measurement Principle and Calibration Procedure for the Continuous Measurement of Carbon Monoxide in the Atmosphere (Non-Dispersive Infrared Spectrometry).

Appendix D--Measurement Principle and Calibration Procedure for the Measurement of Photochemical Oxidants Corrected for Interferences due to Nitrogen Oxides and Sulfur Dioxide.

Appendix E--Reference Method for the Determination of Hydrocarbons Corrected for Methane.

Appendix F--Reference Method for the Determination of Nitrogen Dioxide (24-Hour Sampling Method)

Authority: The provisions of this Part 50 issued under Sec. 4, Public Law 91-604, 84 Stat. 1679 (42 U.S.C. 1857c-4).

Source: The provisions of this Part 50 appear at 36 F.R. 22384, November 25, 1971, unless otherwise noted in the CFR.

PART 51. REQUIREMENTS FOR PREPARATION, ADOPTION, AND SUBMITTAL OF IMPLEMENTATION PLANS

Section

Subpart A--General Provisions

51.1 Definitions.
51.2 Stipulations.
51.3 Classification of regions.
51.4 Public hearings.
51.5 Submittal of plans; preliminary review of plans.
51.6 Revisions.
51.7 Reports.
51.8 Approval of plans.

Subpart B--Plan Content and Requirements

51.10 General requirements.
51.11 Legal authority.
51.12 Control strategy: General.
51.13 Control strategy: Sulfur oxides and particulate matter.
51.14 Control strategy: Carbon monoxide, hydrocarbons, photochemical oxidants, and nitrogen dioxide.
51.15 Compliance schedules.
51.16 Prevention of air pollution emergency episodes.
51.17 Air quality surveillance.
51.17a Air quality monitoring methods.
51.18 Review of new sources and modifications.
51.19 Source surveillance.
51.20 Resources.
51.21 Intergovernmental cooperation.
51.22 Rules and regulations.
51.23 Exceptions.

Part 51 (continued)

Subpart C--Extensions

51.30 Requests for 2-year extension.
51.31 Requests for 18-month extension.
51.32 Requests for 1-year postponement.
51.33 Hearings and appeals relating to requests for one year postponement.
51.34 Variances.

Subpart D--Maintenance of National Standards

51.40 Scope.

AQMA Analysis

51.41 Submittal date.
51.42 Analysis period.
51.43 Guidelines.
51.44 Projection of emissions.
51.45 Allocation of emissions.
51.46 Projection of air quality concentrations.
51.47 Description of data sources.
51.48 Data bases.
51.49 Techniques description.
51.50 Accuracy factors.
51.51 Submittal of calculations.

AQMA Plan

51.52 General
51.53 Demonstration of adequacy.
51.54 Strategies.
51.55 Legal authority.
51.56 Future strategies.
51.57 Future legal authority.
51.58 Intergovernmental cooperation.
51.59 Surveillance.

Part 51 (continued)

51.60 Resources.
51.61 Submittal format.
51.62 Data availability.
51.63 Alternative procedures.

Appendix A--Air Quality Estimation.
Appendix B--Examples of Emission Limitations Attainable with Reasonably Available Technology.
Appendix C--Major Pollutant Sources.
Appendix D--Emissions Inventory Summary (Example Regions).
Appendix E--Point Source Data.
Appendix F--Area Source Data.
Appendix G--Emissions Inventory Summary (other Regions).
Appendix H--Air Quality Data Summary.
Appendix J--Required Hydrocarbon Emission Control as a Function of Photochemical Oxidant Concentrations.
Appendix K--Control Agency Functions.
Appendix L--Example Regulations for Prevention of Air Pollution Emergency Episodes.
Appendix M--Transportation Control Supporting Data Summary.
Appendix N--Emissions Reductions Achievable Through Inspection, Maintenance and Retrofit of Light Duty Vehicles.
Appendix O--[No title--but related to §51.18]
Appendix P--Minimum Emission Monitoring Requirements.
Appendix Q--[Reserved]
Appendix R--Agency Functions for Air Quality Maintenance Area Plans for the _______ AQMA in the State of _________________ for the year __________.

Authority: Part 51 issued under Section 301(a) of the Clean Air Act [42 U.S.C. 1857(a)], as amended by Section 15(c)(2) of Public Law 91-064, 84 Stat. 1713, unless otherwise noted.

Source: Part 51 appears at 36 F.R. 22398, November 25, 1971, unless otherwise noted. AQMA considerations arose from 41 F.R. 18388, May 3, 1976, unless otherwise noted in the CFR. NSR seems to be required by §51.18, with Appendix O intended to assist in developing regulations. Standards are in Part 60.

PART 52. APPROVAL AND PROMULGATION OF IMPLEMENTATION PLANS

Section

Subpart A--General Provisions

52.01 Definitions.
52.02 Introduction.
52.03 Extensions.
52.04 Classification of regions.
52.05 Public availability of emission data.
52.06 Legal authority.
52.07 Control strategies.
52.08 Rules and regulations.
52.09 Compliance schedules.
52.10 Review of new source and modification.
52.11 Prevention of air pollution emergency episodes.
52.12 Source surveillance.
52.13 Air quality surveillance; resources; intergovernmental cooperation.
52.14 State ambient air quality standards.
52.15 Public availability of plans.
52.16 Submission to administrator.
52.17 Severability of provisions.
52.18 Abbreviations.
52.19 Revision of plans by Administrator.
52.20 Attainment dates for national standards.
52.21 Significant deterioration of air quality.
52.22 Maintenance of national standards.
52.23 Violation and enforcement.

Subpart B--Subpart DDD

SIPs for States and Territories

Subpart EEE--Approval and Promulgation of Plans

Appendix A--Interpretive rulings for §52.22(b)--Regulation for review of new or modified indirect sources.

Appendix B-C--[Reserved]

Appendix D--Determination of sulfur dioxide emission from stationary sources by continuous monitors.

Appendix E--Performance specifications and specification test procedures for monitoring systems for effluent stream gas volumetric flow rate.

Authority: 40 U.S.C. 1857c-5, 42 U.S.C. 1857c-5 and 6; 1857g(a); 1859(g).

Source: For Subpart A, 37 FR 10846, May 31, 1972, unless otherwise noted.

PART 60. STANDARDS OF PERFORMANCE FOR NEW STATIONARY SOURCES

Subpart A--General Provisions

Subpart B--Adoption and Submittal of State Plans for Designated Facilities

Subpart C--[Reserved]

Subpart D--Standards of Performance for Fossil-Fuel-Fired Streat Generators

Subpart E--SOP for Incinerators

Subpart F--SOP for Portland Cement Plants

Subpart G--SOP for Nitric Acid Plants

Subpart H--SOP for Sulfuric Acid Plants

Subpart I--SOP for Asphalt Concrete Plants

Subpart J--SOP for Petroleum Refineries

Subpart K--SOP for Storage Vessels for Petroleum Liquids

Subpart L--SOP for Secondary Lead Smelters

Subpart M--SOP for Brass and Bronze Ingot Production Plants

Subpart N--SOP for Iron and Steel Plants

Subpart O--SOP for Sewerage Treatment Plants

Subpart P--SOP for Primary Copper Smelters

Subpart Q--SOP for Primary Zinc Smelters

Subpart R--SOP for Primary Lead Smelters

Subpart S--SOP for Primary Aluminum Reduction Plants

Subpart T--SOP for the Phosphate Fertilizer Industry: Wet Process Phosphoric Acid Plants

Subpart U--SOP for the Phosphate Fertilizer Industry: Superphosphoric Acid Plants

Subpart V--SOP for the Phosphate Fertilizer Industry: Diammonium Phosphate Plants

Subpart W--SOP for the Phosphate Fertilizer Industry: Triple Superphosphate Plants

Subpart X--SOP for the Phosphate Fertilizer Industry: Granular Triple Superphosphate Storage Facilities

Subpart Y--SOP for Coal Preparation Plants

Subpart Z--SOP for Ferroalloy Production Facilities

Subpart AA--SOP for Steel Plants: Electric Arc Furnaces

Part 60 (concluded)

Appendix A--Reference Methods.

Appendix B--Performance Specifications.

Appendix C--Determination of Emission Rate Change.

Appendix D--Required Emission Inventory Information.

Authority: Sections 111 and 114 of the Clean Air Act, as amended by Section 4(a) of Public Law 91-604, 84 Stat. 1678 (42 U.S.C. 1857c-6, 1857c-9).

Source: 36 FR 24877, December 23, 1971, unless otherwise noted in the CFR.

APPENDIX B
SOME SPECIFIC AIR QUALITY MODELS

APPENDIX B
SOME SPECIFIC AIR QUALITY MODELS

In Chapter IV of this report we subdivided air quality simulation models into the following generic categories:

> Rollback
> Isopleth
> Physico-Chemical
  - Grid
  - Trajectory
  - Gaussian
  - Box

In this appendix we associate with each of these generic types a number of specific models. We include many of the models with which we are familiar. Because the list is intended only to be a representative one, we do not enumerate all available models. Many others, particularly Gaussian models, certainly exist and would be appropriate for use in the proper circumstances. In compiling this list, we have drawn heavily from material in Argonne (1977), EPA (1977a), and Roth et al. (1976), as well as various program users' manuals. Also we have made no attempt to screen the models for technical acceptability.

Among the information contained in the accompanying table is the following: model developer, EPA recommendation status, technical description, and model capabilities. The last of these is further subdivided into source type/number, pollutant type, terrain complexity, and spatial/temporal resolution.

TABLE B-1. SOME SPECIFIC AIR QUALITY MODELS

					Capabilities							
					Sources				Resolution			
Category	Name	EPA Recommendation Status	Developer	Description	Number	Type	Pollutant Type	Terrain Complexity	Temporal	Spatial	Form of Output	Problems Addressed
ROLLBACK	Linear Rollback	Accepted by EPA for reactive and non-reactive pollutants; nonverifiable	EPA	A linear relationship is assumed between HC emissions and peak pollutant level	No treatment of individual sources		Oxidant (but has been applied to other reactive and nonreactive pollutants)	Not considered	None (1-hour implied)	Regional only	Percentage cutback required (PCR) in NMHC	Regional oxidant (urban only)
ISOPLETH	EPA EKMA Isopleth Method	Not yet recommended (much active interest, however); nonverifiable	EPA	Isopleths of constant peak O_3 on a plot of NO_x vs. NMHC are constructed using a chemical kinetic mechanism tuned to fit smog chamber data for the isopleth asymptotes. The diagram incorporates diurnal variation in solar radiation; and insolation, dilution, and inversion typical of a stagnant, mid-summer day in LA. Entry to diagram is with 6-9 a.m. NMHC/NO_x ratio and peak observed O_3.	No treatment of individual sources		O_3 NO_x HC	Not considered	Not considered (1-hour implied)	Regional only (urban)	Percentage cutback required (PCR) in NMHC and NO_x	Regional oxidant (urban)
	Whitten	No recommendation status; nonverifiable	Systems Applications, (San Rafael, CA)	Isopleth diagram is similar to one used in EPA method except for a constant sun rather than a diurnally varying one. Entry parameters also differ. Absolute NMHC and NO_x concentrations are used to determine the extent to which the actual airshed resembles a smog chamber.	No treatment of individual sources		O_3 NO_x HC	Not considered	Not considered (1-hour implied)	Regional only (urban)	PCR/NMHC NO_x	Regional oxidant (urban)
PHYSICO-CHEMICAL												
GRID	**Grid-Region Oriented**											
	SAI Urban Airshed Model	No recommendation status; evaluated in several locales	Systems Applications, Incorporated	Three-dimensional grid model based on solution of the atmospheric diffusion equation. Features: time varying emissions and 3-D wind field; 13 chemical species; K-theory diffusivity; a 42-step kinetic mechanism based on carbon-bond chemistry (a lumped mechanism dividing HC into single bonds, fast double bonds, slow double bonds, and carbonyl bonds)—32 steps describe HC, NO_x, and O_3, 6 steps treat the oxidation of SO_2, and 4 steps describe the formation of nitrate, sulfate and organic aerosol; horizontal diffusivity is assumed constant; vertical diffusivity varies in space and time and depends on height, wind speed, surface roughness, and atmospheric stability class.	Any number	Point Area Line Elevated Source	O_3, NO, NO_2, CO, SO_2, HNO_2, H_2O_2, PAN, Total aerosols. Four HC categories (single bond, slow double bond, fast double bond, carbonyl bond)	Terrain features handled thru wind field and surface roughness coefficients	As fine as input data resolution; can handle up to 36-hours	As fine as grid cell size (often 2 km x 2 km)	Spatial concentration maps for each hour for each pollutant of interest; vertical concentration profiles; point predictions at monitoring stations; concentration isopleths	Regional-scale problems; evaluation studies have been carried out for LA Las Vegas, and Denver; with Sacramento and St. Louis soon to follow
	LIRAQ	No recommendation status; initial evaluation in one locale (SF Bay Area)	Lawrence Livermore Laboratory (Livermore, CA)	Two dimensional grid model based on solution of the atmospheric diffusion equation. Each grid cell is bounded by the terrain and the inversion base. Pollutants are assumed well mixed. An empirical algorithm relates cell-averaged concentrations to ground-level concentrations. Features: time varying emissions and 2D winds; 15 chemical species; a 51-step kinetic mechanism (a lumped HC mechanism dividing HC into "olefins and highly reactive aromatics"; "paraffins, less reactive aromatics, and some oxygenates"; "aldehydes, some aromatics, and ketones"); only horizontal diffusivity is considered; mass consistent wind field.	Any number	Point Area Line	O_3, NO, NO_2, CO, HNO_2, H_2O_2, HNO_3, N_2O_5, NO_3, HO_2, RCO_3, RNO_2. Three HC categories	Horizontal features can be handled through wind field; vertical features thru cell vertical dimension	As fine as input data resolution (Time scale up to 24-hours)	As fine as grid cell size	Roughly the same as for the SAI model	Regional-scale problems; an evaluation study has been conducted for the SF Bay Area.

TABLE B-1 (Continued)

Category	Name	EPA Recommendation Status	Developer	Description	Capabilities						Form of Output	Problems Addressed
					Sources		Pollutant Type	Terrain Complexity	Resolution			
					Number	Type			Temporal	Spatial		
	PICK--Sklarew et al.	No recommendation status; evaluated in LA Basin	Systems, Science, and Software, Inc. (La Jolla, CA)	Three-dimensional grid model. The particle-in-cell technique is used to solve the atmospheric diffusion equation. The motion of each particle is influenced by winds and pseudo-diffusion velocity. Sources and chemical reactions are accounted for through creation, destruction, or change of mass of the particles. Information as to location and mass of each particle must be maintained. Features: time varying emissions and 3-D winds; five chemical species based on 12-step mechanism developed by Eschenroeder and Martinez; diffusivities are assumed constant.	Any number	Point Area Line	O_3, NO, NO_2, CO, RHC	Can handle some terrain complexity through wind-field input	As fine as input data resolution (Time scale hours)	As fine as grid cell size	Roughly the same as for the SAI model	Regional-scale problems; an evaluation study has been conducted for the LA Basin (agreement with observation not too good for 3 stations reported in 1971 paper by Sklarew et al.)
	Grid-Specific Source Oriented											
	EGAMA--Egan and Mahoney	No recommendation status	Environmental Research and Technology	Simulates the dispersion mechanisms of grid-cell emissions using numerical solution of the basic tracer equation (mass conservation for a species in a planer non-divergent flow field). Can be adapted to source specific applications such as near field dispersion for highway sources, fumigation near to point sources, and long-range transport. Features: time varying emissions and 2-D or pseudo-3-D winds; two species chemistry/decay; no plume rise formula; emissions assumed well mixed; time varying vertical diffusivity; horizontal diffusivity often neglected; K-theory.	Few	Highway Point	CO, SO_x	Some complexity	As fine as input data resolution	As fine as grid cell size (can be used for some micro-scale applications)	Concentration fields at grid locations.	Highway impact studies (2-D version); limited use of 3-D version in long-range SO_x transport; fumigation; downwash.
	DEPICT--Detailed Examination of Plume Impact in Complex Terrain	No recommendation status	Science Applications, Inc. (Sklarew et al.)	A 3-D grid-based point source model. It is applicable to point sources in rural environments. Solves conservation of mass equation. It calculates its own 3-D wind field using potential flow incorporating temperature profiles or stability classes. Incompressible gas assumed. Winds are projected upward from the point measured based on a power law. Features: Briggs plume rise; horizontal diffusivity a function of stability class; vertical diffusivity based on Smith and Howard algorithm (somewhat like K-theory); uses Eschenroeder 16-step chemical mechanism or Hecht-Seinfeld-Dodge 39-step mechanism; time varying emissions and winds input.	Up to 10	Point	SO_2, NO, NO_2, O_3	Complex	As fine as input data resolution	As fine as grid cell size	Hourly spatial concentration field at grid locations	Point source impact studies; power plants, smelters, for example; specific examples are Garfield smelter (SF_6), Navajo Generating Station (SO_2), and Ormond Beach Generating Station (SF_6, NO, NO_2, O_3)
TRAJECTORY	Trajectory-Region Oriented											
	DIFKIN--Eschenroeder and Martinez	No recommendation status	General Research Corporation--GRC (Santa Barbara, CA); now offered by ER&T	A column or parcel of air moving under the influence of local meteorological conditions is allowed to traverse a 2-D grid. Emissions are injected into the parcel and reactions occur along the track. Features: time varying emissions and 2-D local winds are input; five species are simulated using a 12-step reaction mechanism; the vertical column is subdivided into several cells; vertical diffusivity is a function of temperature gradient and height above surface; no horizontal diffusivity.	Any number	Point Line Area Elevated Source	O_3, NO, NO_2, CO, RHC	Not explicit (but horizontal features can be handled thru the wind field)	As fine as input data resolution	Only along trajectory track; several could be run side-by-side, however	Temporal concentration history in air parcel	Regional oxidant; applied to LA Basin (six days in 1969)
	REM--Wayne et al.	No recommendation status	Pacific Environmental Services, Inc. (Santa Monica, CA)	Similar to DIFKIN except that the air column is treated as a single well-mixed cell. Features: 11 species are simulated using a 32-step mechanism designed to treat propylene and less-reactive HC; no vertical diffusivity needed; emissions are converted to equivalent propylene and LRHC	Any number	Point Line Area	O_3, NO, NO_2, CO, HNO_3, C_3H_6, CH_3CHO, HCHO, CH_3ONO_2, $C_2H_5NO_3$, LRHC	Not explicit (but horizontal features can be handled thru the wind field)	As fine as input data resolution	Only along trajectory track; several could be run side-by-side	Temporal concentration history in air parcel	Regional oxidant; applied to LA Basin (six days in 1969)

TABLE B-1 (Continued)

Category	Name	EPA Recommendation Status	Developer	Description	Capabilities						Form of Output	Problems Addressed
					Sources		Pollutant Type	Terrain Complexity	Resolution			
					Number	Type			Temporal	Spatial		
	ARTSIM	No recommendation status	ER&T	The model is trajectory oriented and intended to be used for regional application. It appears to be similar to DIFKIN in that the air column allows up to 10 vertical layers. Features: hourly emissions and horizontal 2-D winds are input; simulated species include four HC classes (alkenes, alkanes, aromatics, and aldehydes) as well as oxidants, SO_2, and sulfate' a 54-step mechanism is employed; no horizontal diffusion; vertical diffusivity specified at up to 10 vertical levels with time variation.	Any number	Point Area Line	O_3, NO, NO_2, SO_2, Sulfate Four HC groups (alkenes, alkanes, aromatics, aldehydes)	Not explicit but horizontal feature can be handled thru wind field	As fine as input data resolution	Only along trajectory track; several could be run side-by-side	Temporal concentration history in air parcel	Regional oxidant; applied to Las Vegas (1976), Truckee (1976) and SF Bay Area (1974) as well as LA Basin (1972-Eschenroeder and Martinez)
	Trajectory-Specific Source Oriented											
	RPM	No recommendation status	Systems Applications (for the California Air Resources Board--CARB)	The model is designed to estimate concentrations of reactive species downwind of a single point or areal source. Based on Lagrangian (moving-with-air-parcel) version of mass conservation equation, allowing for background entrainment, the air parcel containing the emitted pollutants is allowed to drift downwind. The parcel expands from the plume height according to measured plume width and depth as functions of downwind distance or to the Pasquill-Gifford methods. Features a modified H-S-D mechanism for HC-NO_x-SO_2; 2-D wind field; plume rise input.	Single source	Point Areal	O_3, NO, NO_2, SO_2, Sulfate	No terrain interaction currently	As fine as input data resolution, long-range transport as well	Resolution all the way to source (near-, medium-, and far-field)	Temporal concentration history in downwind direction	Single source problems, e.g. refineries, power plants for fumigation; trapping; applied to: Moss Landing PP, Monterey; Los Alamitos PP, LA; Haynes PP, LA; Mobile Oil R, LA; Four Corners PP, Farmington, NM; Hobbs PP, Hobbs, NM; Jefferson PP, Jefferson, Texas
	LAPS	No recommendation status	ER&T	The model is designed to calculate concentration fields downwind of single or multiple concentrated sources. The air parcel is allowed to drift downwind, dispersing laterally and vertically. Features: equilibrium coupling of NO, NO_2, and O_3; first order conversion of SO_2 to sulfate; eddy diffusivities; 2-D wind field; Briggs plume rise; up to 7 species can be specified.	Up to 10 point sources; separate areal sources	Point, areal, elevated sources	O_3, NO, NO_2, SO_2, sulfate	No terrain interaction	As fine as input data resolution	Resolution all the way to source	Vertical concentration maps (10 vert. x 20 hours sta.); ground concen. maps and contours; concen. vs. dist.; ground concen. crossplot	Single or few sources problems; only analytical problems attempted, e.g., steady-state Gaussian plumes
GAUSSIAN	Long-Term Averaging											
	AQDM--Air Quality Display Model	Recommended by EPA in guidelines (No. 26)	TRW (for Public Health Service)	This is a climatological steady state Gaussian plume model that estimates the annual arithmetic average SO_2 and particulate concentration at ground level. A statistical model based on Larson (1969) is used to transform the average concentration data from a limited number of receptors into an expected geometric mean and maximum concentration values for several averaging times. Features: treats one or two pollutants simultaneously; Holland (1953) plume rise; no plume rise for areal sources; no temporal variation in sources; 16 wind directions; 6 wind speed classes; 5 stability classes (Turner, 1964); Pasquill-Gifford stability coefficients; no chemical mechanism; perfect reflection at ground; no effect at mixing height until $\sigma_z \geq 0.47H$ (when $x = x_2$); for $x > x_2$, uniform mixing; no variation in wind speed with height; linear superposition of sources; $\sigma_z(x) = ax^b + c$; does not treat fumigation or downwash; Larson procedure assumes log-normal concentration distribution and power law dependence of median and maximum concentrations or averaging time.	Many (up to 12 user-specified receptor locations; up to 225 receptors located on a uniform rectangular grid)	Point, areal, elevated sources	SO_2, TSP (could be used for NO_x with NO_2 obtained thru use of an appropriate factor)	Relatively flat terrain; no height difference allowed between source and receptors	Steady-state; averaging time = 1 mo. to 1 yr.; Larsen procedure can be used to transform to 1-24 hour averages	Regional scale	1 mo.-1 yr. averaged concentrations; individual point, area source culpability list for each receptor	Regional long-term averages for relatively inert pollutants; urban areas primarily

TABLE B-1 (Continued)

Category	Name	EPA Recommendation Status	Developer	Description	[illegible]						Form of Output	Problems Addressed
					Sources		Pollutant Type	Terrain Complexity	Resolution			
					Number	Type			Temporal	Spatial		
	CDM and CDMQC--Climatological Display Model	Recommended by EPA in guidelines (No. 27)	EPA	This is a climatological steady state Gaussian plume model for determining long-term (seasonal or annual) arithmetic average concentrations at any ground level receptor in an urban area. Useful for relatively inert pollutants. Features: Larsen procedure; $\sigma_z(x) = ax^b$; treats one or two pollutants simultaneously; 16 wind directions and 6 speed classes; no plume rise for areal sources; Briggs (1971) plume rise; no chemical reactions; power law correction on elevated wind; 2-D wind; day/night variation in emissions, the same factor for all sources; 5 stability classes (Turner, 1974); exponential decay for physical removal; perfect reflection at ground; no effect vertically until $\sigma_z(x) = 0.8H$ (where H = mixing height); uniform mixing beyond that point; dispersion coefficient from Turner; linear superposition of sources.	Many	Point, areal, elevated sources	SO_2, TSP (could be used for NO_x with NO_2 obtained thru use of an appropriate factor	Relatively flat terrain; no height difference allowed between source and receptors	Steady-state; averaging time = 1 mo. to 1 yr.; Larsen procedure can be used to transform to 1-24 hour averages	Regional scale	1 mo. to 1 yr. averaged concentrations; source-receptor culpability list (CDMQC only)	Regional long-term averages for relatively inert pollutants; urban areas primarily
	TCM--Texas Climatological Model	No recommendation status	Texas Air Control Board	This is a climatological steady state Gaussian plume model similar to CDM but incorporating design features "reducing run time by as much as two orders of magnitude." Features: downwash and fumigation not considered; all sources have a single average emissions rate for the averaging period (i.e., month, season, year); Pasquill-Gifford-Turner stability classes; mixing height not a factor because no effect for typical climatology.	Unlimited (arbitrary receptor locations--max 50x50)	Point, line, areal, elevated sources. "tall stack" sources in short-term "sequential" mode	SO_2, TSP, CO, NO_x	Relatively flat terrain; no height difference allowed between source and receptors	Steady-state; averaging time = 1 mo. to 1 yr.; Larsen procedure can be used to transform to 1-24 hour averages	Regional scale	Mean concentration; concentration at grid points (up to 50x50); a listing of the 5 highest contributors to concentrations at each grid point	Regional long-term averages for relatively inert pollutants; urban areas primarily
	ERTAQ--can also be used for short-term averaging	No recommendation status	ERT	This is a steady-state sector-averaged Gaussian plume model that calculates concentrations of up to six pollutants from an unlimited number of point, line, and areal sources. The model can be operated either in the "climatological" mode or the "sequential" mode for short-term averaging times. Features: crosswind dispersion function may be sector-averaged over 22.5°; for "sequential" mode and "tall stacks," the crosswind dispersion function is given by the expected value within the 22.5° sector for receptors within the downwind sector; for receptors adjacent to the downwind sector, a formulation is used which avoids centerline one-hour values when accumulating concentration estimates for multiple-hour averages; Briggs plume rise; stack tip downwash (Gifford) for tall stacks; wind speed power law; half-life decay factors for species; chemistry not treated directly; perfect reflection at ground and mixing layer; unique emissions rate for each source that may be varied diurnally, weekly or monthly; 5 stability classes.	Unlimited (up to 128 receptor points at any selected locations)	Point line areal elevated sources "tall stack" sources in short-term "sequential" mode	SO_2, TSP, CO, NO_x	Flat and hilly terrain; a "tall stack" terrain correction is available for "sequential" mode but not "climatological" mode; also a unique elevation can be specified for receptors; plume and mixing depth respond to terrain obstacles	Steady-state; can handle short-term in "sequential" mode (1, 3, 8, and 24 hr) and long-term in "climatological" mode (1 mo., seasonal, 1 yr.)	Regional scale	Concentrations at each receptor	Regional long- and short-term averages for relatively inert pollutants; urban and rural areas

TABLE B-1 (Continued)

					Capabilities							
					Sources				Resolution			
Category	Name	EPA Recommendation Status	Developer	Description	Number	Type	Pollutant Type	Terrain Complexity	Temporal	Spatial	Form of Output	Problems Addressed
	Short-Term Averaging											
	APRAC-1A	Recommended by EPA in guidelines (No. 34 and 35)	EPA (developed by Stanford Research Institute)	This is a model which calculates hourly average CO concentrations for urban areas. Contribution from dispersion on 3 scales are calculated: extraurban, mainly from sources upwind of city of interest (simulated using a box model); intraurban, from freeway, arterials, and feeders (Gaussian plume until it equals box model value which is used thereafter); and local, from street canyon effects. Features: no plume rise, fumigation or downwash; helical circulation in street canyons; hourly varying traffic emissions and 2-D wind field; $\sigma_z(x) = ax^b$; link emissions are aggregated into area sources; no wind power law; 6 stability classes (Turner); dispersion coefficients from McElroy and Pooler (1968), modified using Leighton and Ditmar (1953); no chemistry; perfect reflection at surface and inversion (ignores latter until concentration equals that calculated using box model and uses that thereafter).	Many (an extensive traffic inventory is required); up to 10 receptors; 4 internally defined receptors are used on each street where street canyon effects are considered	Line	CO, TSP	Relatively flat terrain; street canyon effects can be handled	Steady-state; hourly averages	Regional	Hourly concentration values at each receptor; frequency distribution using hourly values	Regional CO problems from traffic sources; urban areas
	CRSTER—this also can be used for annual averaging	Recommended by EPA in guidelines (No. 13)	EPA	Steady-state Gaussian plume model applicable in uneven terrain. Features: 7 stability classes (Turner, Pasquill); dispersion coefficients from Turner; no chemistry; Briggs plume rise; no fumigation or downwash; perfect reflection at surface and inversion (multiple reflections until $\sigma_z = 1.6H$ and uniform thereafter); mixing height is constant and follows topography; monthly emissions variations.	Single source up to 19 stacks (all assumed at the same location; unique topographic height for each receptor	Point	CO, SO_2, NO_x, TSP	Some application in uneven terrain	Steady-state; (1 and 24-hour) averages; also annual averages	Near source and downwind	Highest and second highest 1 hr. and 24 hr. concentrations at each receptor for year plus annual arith. average; concentration for each day; frequency distribution data	Point source complex terrain problems; power plants; rural areas (isolated source)
	HANNA-GIFFORD	Recommended by EPA in guidelines (Nos. 28 and 29)	EPA	Steady-state Gaussian plume model used to calculate dispersion from urban area sources. Analytic integration of area sources. All sources upwind of each receptor area are summed. It is most applicable in areas where no point source information is available. Features: perfect reflection at ground; mixing height reflection not considered; hourly emissions and winds; $\sigma_z(x) = ax^b$; dispersion coefficients from Smith (1968); stability classes from Smith; narrow plume approx. (no horizontal dispersion); no plume rise; no chemistry.	Many	Area	SO_2, TSP, CO	Simple	1-hr., 8-hr. and 24-hr. averages (although an annual average can be estimated	Regional	Hourly concentration values at receptors	Regional problems involving inert pollutants; urban area
	HIWAY	Recommended by EPA in guidelines (No. 36)	EPA	Steady-state (S-S) Gaussian plume model that computes the hourly concentrations of nonreactive pollutants downwind of roadways. Based on analytic integration of line source. It is applied to each lane of traffic. Features: no chemistry; perfect reflection at surface and inversion; one road or highway segment per run; 6 stability classes (Turner); dispersion coefficients from Turner; for distances < 100 m, coefficient from Zimmerman and Thompson (1975); no wind power law; hourly emissions and 2-D wind.	Up to 24 (arbitrary receptor and release heights)	Line	CO, TSP	Level terrain	Hourly (1-24 hr. average)	Near to medium field downwind	One-hr. average concentrations at each receptor	Regional- or highway-specific problems for nonreactive pollutants
	PTMTP	Recommended by EPA in guidelines (No. 17)	EPA	An S-S Gaussian plume model that considers multiple point sources. It is based on linear additivity of individual source effects. Features: hourly emissions and winds; Briggs plume rise; no fumigation or downwash; no wind power law; Turner stability classes and dispersion coefficients (horizontal and vertical); no chemistry; perfect reflection at surface and inversion (multiple reflection).	Up to 25 (up to 30 receptors)	Point (elevated)	SO_2, TSP	Flat terrain	Hourly (1-24 hr. average)	Regional	Hourly concentrations; source contribution list at each receptor; average concentrations	Regional point source problems; urban area; nonreactive pollutants

TABLE B-1 (Continued)

Category	Name	EPA Recommendation Status	Developer	Description	Capabilities						Form of Output	Problems Addressed
					Sources		Pollutant Type	Terrain Complexity	Resolution			
					Number	Type			Temporal	Spatial		
	PTDIS	Recommended by EPA	EPA	A steady-state Gaussian plume model that estimates short-term center-line concentrations directly downwind of a point source. Features: same as PTMTP.	Single source (up to 50 receptors, all at ground level)	Point (elevated)	SO_2, TSP	Flat terrain	Hourly (1-24 hr. average)	Near to medium field downwind	Centerline, ground-level concentrations; isopleth halfwidths of up to 8 concentration levels	Single source problems; nonreactive pollutants
	PTMAX	Recommended by EPA	EPA	An S-S Gaussian plume model that finds the maximum short-term concentrations from a single point source as a function of stability and wind speed. Features: same as PTMTP except inversion reflection not considered.	Single source	Point (elevated)	SO_2, TSP	Flat terrain	Hourly (1-24 hr. average) and less than 1 hr.	Point at which max. occurs; near to medium field	Maximum ground-level concentrations; distance to maximum for all stability classes and wind speeds	Single source problem; nonreactive pollutants
	RAM	Recommended by EPA in guidelines (No. 32)	EPA	An S-S Gaussian plume model for averaging times from 1-hour to 24-hours and use in level or gently rolling terrain. Features: no chemistry; Briggs plume rise; no fumigation downwash; hourly emissions and 2-D winds; wind power law (function of stability class); 3 possible release heights for area sources; exponential decay; stability class determined internally by Turner (1964); 6 classes; dispersion coefficients from McElroy and Pooler (1968)--urban--and Turner (1969)--rural; perfect reflection from surface and inversion; multiple reflection until $\sigma_z(x) = 1.6H$ and uniform mixing thereafter; mixing height determined from twice-daily temperature soundings (stability class, also).	Many (receptors are all at the same height)	Point areal	SO_2, TSP	Flat terrain	Hourly and averaged up to 24 hrs.	Regional (urban) and rural	Hourly and average concentrations at receptors; limited source contribution list; cumulative frequency distribution data	Regional problems for nonreactive pollutants; urban and rural areas
	VALLEY—This can also be used for annual averaging (climatological mode)	Recommended by EPA in guidelines (No. 14)	EPA	An S-S Gaussian plume model for calculating annual and maximum 24-hour average SO_2 and TSP from single point sources in complex terrain. Features: climatological and short-term modes; 16 wind directions and 6 wind speed categories; Briggs plume rise (1971, 1972); 5 stability (urban) classes (Turner, 1964); dispersion from Pasquill (1961) and Gifford (1961); 6 stability classes for rural; no wind power law; exponential decay for chemistry and removal.	Up to 50 (112 receptors on radial grid; can be at different topological hgts.)	Point, areal (treated as a point source)	SO_2, TSP	Complex terrain	Short- and long-term average (24-hr. and annual)	Regional (urban and rural)	Short-term mode; second highest 24-hr. concentration, source contribution list; long-term mode: arithmetic means, source contribution list	Point source problems for nonreactive pollutants in complex terrain; rural and urban areas
	TEM—Texas Episodic Model	No recommendation status	Texas Air Control Board	An S-S Gaussian plume model for predicting short-term concentrations (10 min. to 24 hour) from multiple point and area sources. Calculations are performed for 1 to 24 scenarios (meteorology, averaging time, and mixing height). Features: Briggs plume rise; mixing height penetration factor; up to 3 pollutants; no chemistry but exponential decay; no downwash on fumigation; wind power law; Pasquill-Gifford-Turner stability classes dispersion coefficient from Turner; perfect reflection from surface and inversion until $\sigma_z = 0.47H$	Up to 300 point sources; up to 200 areal sources (up to 50x50 receptor grid)	Point, areal	SO_2, TSP	Flat terrain	Short-term (1, 3, 24 hrs.)	Regional (urban)	Mean concentration for each grid point (10 min., 30 min., 1 hr., 3 hrs., and 24 hrs.); printed plot; culpability list	Regional problems for nonreactive pollutants; urban area

TABLE B-1 (Concluded)

Category	Name	EPA Recommendation Status	Developer	Description	Capabilities: Sources: Number	Capabilities: Sources: Type	Capabilities: Pollutant Type	Capabilities: Terrain Complexity	Capabilities: Resolution: Temporal	Capabilities: Resolution: Spatial	Form of Output	Problems Addressed
	TAPAS--Topographic Air Pollution Analysis System	No recommendation status	USDA Forest Service	This model combines a simulation of the wind field over mountainous terrain with a Gaussian derived diffusion model. It provides an estimate of the total allowable emissions within each of a number of grid cells (ranging from 0.25 km^2 to 9 km^2) to maintain a preselected level of air quality. The diffusion model is employed in each grid cell to provide an estimate of the mixing conditions within these cells. These conditions are combined with the Pollutant Standards Index such that a maximum allowable emission is calculated. Features: wind model (Cressman objective analysis, potential flow over topography, influences of surface temperature and roughness); Gaussian model (σ_y and σ_z from Turner, effects of mass flow divergence included, stability classes from Turner, no upper bound on diffusion although the wind is calculated assuming a lid at a specified height above the topography); the calculated wind follows the terrain and thus gives a vertical wind component; no chemistry; no explicit treatment of plume behavior.	Many	Point (no distinction made between point, line, and areal sources	SO_2, TSP, CO	Complex	Both short-term and long-term estimates	Limited regional	Allowable emissions in each grid cell for each pollutant of interest	Limited regional impact problems in complex terrain; nonreactive pollutants
	AQSTM--Air Quality Short Term Model	No recommendation status	Illinois Environmental Protection Agency	An S-S Gaussian plume model for estimating short term concentration averages from multiple point sources in level or complex terrains. It can simulate late inversion breakup fumigation, lake shore fumigation, and atmospheric trapping. Features: one or two pollutants simultaneously; no chemistry; Briggs plume rise; no downwash; wind power law; user-supplied stability classes; dispersion coefficients from Turner (1969); perfect reflection at ground and mixing height.	Up to 200 sources (up to 900 receptors located on a uniform rectangular grid); unique topographic elevation for both	Point (elevated)	SO_2, TSP	Mostly flat terrain, but some corrections for complex terrain	Short-term (1, 3, and 24 hr. averaging)	Regional	Average concentrations at receptors; source contributions at receptors	Regional point source problems for nonreactive pollutants; urban areas; shorelines-
	CALINE-2	No recommendation status	California Air Resources Board (CARB)	An S-S Gaussian line source model for traffic impact assessment. Features: no chemistry; perfect ground reflection; Pasquill stability classes; hourly emissions; some accounting for depressed highways	Many (an extensive traffic inventory is required)	Line	CO	Relatively flat terrain	Short-term		Hourly concentrations at receptors	Regional CO problems from traffic sources
BOX	ATDL--Hanna	No recommendation status	Atmospheric Turbulence and Diffusion Laboratory--ATDL (Oak Ridge, Tenn.)	The region of interest is assumed to be encompassed by a single cell or box, bounded by the inversion above and the terrain below. All concentrations are assumed to be in steady-state. Features: for given time, constant emissions rate and simple winds; seven-step chemical mechanism proposed by Friedlander and Seinfeld (1969); uniform and constant wind and constant mixing depth.	All emitting into a single box		O_3, NO, NO_2, RHC	Not explicit	Temporal resolution can be obtained by varying initial conditions to match a temporal pattern	No resolution	Concentration values at the time considered	Regional oxidant; it was applied to LA Basin (30 Sept. 1969 data). Ozone predictions were low.

APPENDIX C

SOME SPECIFIC MODEL PERFORMANCE MEASURES

APPENDIX C

SOME SPECIFIC MODEL PERFORMANCE MEASURES

Having discussed model performance measures in generic terms in Chapter V, we now present some specific examples. We discuss each of the four generic types of performance measures: peak, station, area, and exposure/dosage. We include scalar, statistical, and "pattern recognition" variants.

1. PEAK PERFORMANCE MEASURES

The use of a performance measure of this type requires the modeler to know information about both the predicted and the "true" concentration peak. The measurement network must be so situated as to insure a high probability of sensing the "true" peak concentration or a value near to it. There are three characterizing parameters of interest: peak concentration level, spatial location, and time of occurrence. The predicted and observed values of some or all of these may be available for comparison. Differences in their predicted and observed values represent the performance measures of interest. These peak measures are summarized in Table C-1.

Each measure conveys separate but related information about model behavior in predicting the concentration peak. Their values should be examined in combinations. Several combinations of interest and some of their possible interpretations are shown in Table C-2. The table is not intended to include all combinations and interpretations. Rather, it illustrates by example how inferences can be made about model performance through the joint use of performance measures.

TABLE C-1. SOME PEAK PERFORMANCE MEASURES

Type	Performance Measure
Scalar	a. Difference* in the peak ground-level concentration values.
	b. Difference in the spatial location of the peak.
	c. Difference in the time at which the peak occurs.
	d. Difference in the peak concentration levels at the time of the observed peak.
	e. Difference in the spatial location of the peak at the time of the observed peak.
Pattern recognition	Map showing the locations and values of the predicted maximum one-hour-average concentrations for each hour.

---

[*] "Difference" as used here usually refers to "prediction minus observation."

Several points are contained in Table C-2. While a large difference in peak concentration levels might in itself be sufficient reason to question a model's performance, a simple difference in peak location might not. If the concentration residual (the difference between predicted and observed values) at the peak is small (good agreement) and yet there is a difference in the spatial location of the peak, this may be due mostly to slight errors in the wind field input to the model. The slight offset in the location of the peak might cause predicted and measured concentrations to disagree at specific monitoring stations, particularly if concentration gradients within the pollutant cloud are "steep." However, a small displacement in the concentration field, unless it resulted in a large change in population exposure and dosage, may not be a serious problem. Model performance might be otherwise acceptable.

TABLE C-2. SEVERAL PEAK MEASURE COMBINATIONS OF INTEREST AND SOME POSSIBLE INTERPRETATIONS

Residual Values			
Concentration Level	Location	Timing	Some Possible Interpretations
Event-Related*			
Small	Small	Small	Model performance in predicting the concentration peak is acceptable
	Large	Small	Model performance is still good in predicting the peak concentration level There is a possible error in the wind field input
		Large	Concentration level prediction is good There is a possible error in wind field input There is a possible error in the chemistry package or emissions input
Large	Any value	Any value	Model performance is probably unacceptable
Fixed-Time†			
Large	Large	--	Model performance may or may not be acceptable; event-related (peak) residuals must be examined to make a final judgment
	Small	--	Model performance is probably unacceptable Pollutant transport is handled acceptably well There is a possible error in the chemistry package, the emissions input, or the inversion height time and spatial history

---

* Residual values are calculated at the time an event occurs (the peak).

† Residual values are calculated at a fixed time (the time of the observed peak).

On the other hand, if the spatial offset of the location of the peak is accompanied by a significant difference between the predicted and observed times at which the peak occurs, more serious problems might be suspected. Not only might there be a wind field problem, but the chemical kinetic mechanism may be giving erroneous results (if the pollutant species of interest is a reactive one). Alternatively (or additionally), one might suspect that the emissions supplied as input to the model were not the same as those injected into the actual atmosphere. Another possibility also exists. Slight differences between the modeled and actual wind field might result in the air parcel in which the peak occurs following a space-time track having sufficiently different emissions to account for differences in peak concentration values.

Additional clarity of interpretation can be achieved in another way. We can compare concentration level, location and timing, not just at the time a specific event occurs (the peak, for instance) but also at a fixed time (the time at which the observed peak occurs, for example). Suppose that the concentration level residual at that fixed time (the difference between maximum predicted concentration and the observed peak value) is large but the spatial one is not. In this case, one could conclude that the model reproduced the pollutant transport process but was unable to predict concentration levels. This could result from many causes, among which are errors in the chemical kinetic mechanism, the emissions input, or the inversion height space/time profile. Whatever the cause, however, the conclusion remains the same: Model performance is probably inadequate.

Alternatively, if both the fixed-time concentration level and location residuals are both large, a firm conclusion about model acceptability may be premature. Performance may or may not be satisfactory. A comparison with the event-related peak performance measures is necessary before a final judgment is made.

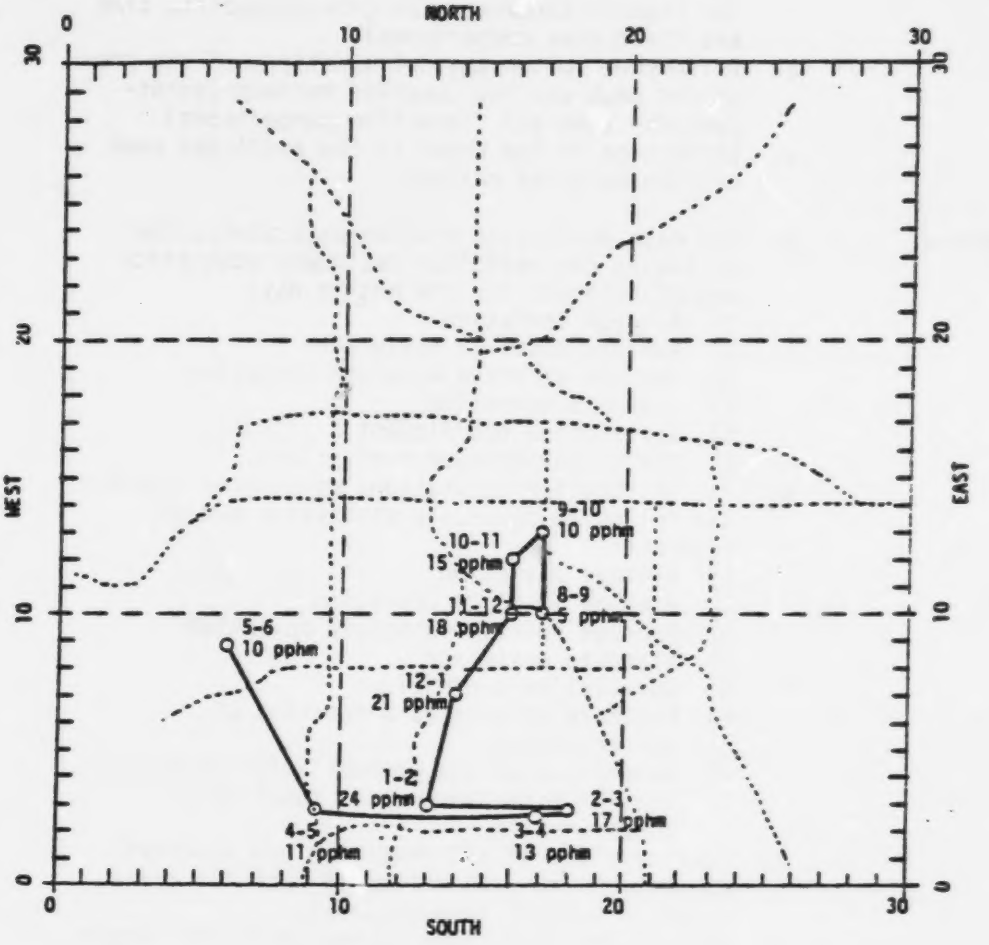
If the model being used is capable of sufficient spatial and temporal resolution, a "pattern recognition" performance measure may be of some use: a map showing the locations and values of the predicted maximum concentrations at several times during the day. Such a map is shown in Figure C-1. It was produced using the SAI Urban Airshed Model simulating conditions in the Denver Metropolitan region.

2. STATION PERFORMANCE MEASURES

The use of a station performance measure requires the modeler to know, usually at each hour during the daylight hours, the values of both the predicted and observed concentrations at each monitoring stations. From the two concentration time histories at each site, a number of performance measures are listed in Table C-3, divided into three categories: scalar, statistical, and "pattern recognition."

Station measures are the performance measures whose use is most feasible in practice. Their calculation is based upon the comparison of model predictions with observational data in the form that it is most often available--a set of station measurements. By contrast, peak measures require the observation of the "true" peak. If this peak value is not the same as the value recorded at that station in the monitoring network measuring the highest level, if the location of the peak is somewhere other than at that station, and if its time of occurrence is different than the time of the peak observation, then the calculation of peak performance measures may not be feasible. Although one can sometimes use numerical methods to infer from station data the level, location and timing of the peak, results are subject to uncertainty.

Similarly, area and exposure/dosage measures require knowledge of the "true" spatially and temporally varying concentration field. However, unless circumstances are simple and the monitoring network is exceptionally extensive and well-designed, the "true" concentration field will not be known. The only data available will consist of station measurements. Inference of the concentration field from such data can often be an uncertain and error prone process.


NORTH
SOUTH
WEST
EAST
0
10
20
30
9-10
10 pphm
10-11
15 pphm
11-12
18 pphm
8-9
5 pphm
5-6
10 pphm
12-1
21 pphm
1-2
24 pphm
2-3
17 pphm
3-4
13 pphm
4-5
11 pphm

Meteorology of 3 August 1976

FIGURE C-1. LOCATIONS AND VALUES OF PREDICTED MAXIMUM ONE-HOUR-AVERAGE OZONE CONCENTRATIONS FOR EACH HOUR FROM 8 a.m. TO 6 p.m.

TABLE C-3. SOME STATION PERFORMANCE MEASURES

Type	Performance Measure
Scalar	a. Concentration residual at the station measuring the highest concentration (event-specific time and fixed-time comparisons).
	b. Difference in the spatial locations of the predicted peak and the observed maximum (event-specific time and fixed-time comparisons).
	c. Difference in the times of the predicted peak and the observed maximum.
Statistical	a. For each monitoring station separately, the following concentration residuals statistics are of interest for the entire day: 1) Average deviation 2) Average absolute deviation 3) Average relative absolute deviation 4) Standard deviation 5) Correlation coefficient 6) Offset-correlation coefficient.
	b. For all monitoring stations considered together, the following residuals statistics are of interest: 1) Average deviation 2) Average absolute deviation 3) Average relative absolute deviation 4) Standard deviation 5) Correlation coefficient 6) Estimate of bias as a function of concentration 7) Comparison of the probabilities of concentration exceedances as a function of concentration
	c. Scatter plots of all predicted and observed concentrations with a line of best fit determined in a least squares sense.
	d. Plot of the deviations of the predicted versus observed points from the perfect correlation line compared with estimates of instrumentation errors.
Pattern recognition	a. Time history for the modeling day of the predicted and observed concentrations at each site.
	b. Time history of the variations over all stations of the predicted and observed average concentrations.
	c. At the time of the peak (event-related), the ratio of the normalized residual at the station having the highest value to the average of the normalized residuals at the other stations.

a. Scalar Station Performance Measures

Since the "true" concentration peak is not always known with confidence, a surrogate is needed for determining model performance in predicting the concentration peak. Such a measure is often based upon a comparison of the predicted and observed concentrations at the station measuring the highest value during the day. The comparison can be done at an event-related time (the peak) or a fixed time. Since the values of the measures may differ at the two times, the implications of those differences should be considered carefully.

b. Statistical Station Performance Measures

Many statistical station performance measures are of use. Sometimes the behavior of the concentration residuals at a single station is considered. At other times, the overall behavior of the residuals averaged over all stations is the focus of interest. In either case, however, several of the statistical performance measures remain the same. We define them here (the tilde ~ denotes "predicted," while m is the pollutant species, n is the hour of the day, k is the station index, K is the number of stations being considered, and N is the number of hours being compared:

- Average Deviation

$$\mu^{m} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\tilde{C}_{k}^{m,n} - C_{k}^{m,n} \right) \tag{C-1}$$

- Average Absolute Deviation

$$|\mu|^{m} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \left| \tilde{C}_{k}^{m,n} - C_{k}^{m,n} \right| \tag{C-2}$$

- Average Relative Absolute Deviation

$$|\bar{\mu}|^{m} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\left| \tilde{C}_{k}^{m,n} - C_{k}^{m,n} \right|}{C_{k}^{m,n}} \tag{C-3}$$

> Standard Deviation

$$(\sigma^{m})^{2} = \frac{1}{KN - 1} \sum_{n=1}^{N} \sum_{k=1}^{K} \left[\left(\tilde{C}_{k}^{m,n} - C_{k}^{m,n} \right) - \mu^{m} \right]^{2} \qquad \text{(C-4)}$$

or, alternatively,

$$(\sigma^{m})^{2} = \frac{1}{KN - 1} \left\{ \sum_{n=1}^{N} \left[\sum_{k=1}^{K} \left(\tilde{C}_{k}^{m,n} - C_{k}^{m,n} \right) \right]^{2} - N(\mu^{m})^{2} \right\} \qquad \text{(C-5)}$$

The first three of these relations are designed to measure the mean difference between predicted and observed concentration, either at a particular station (K = 1) or averaged over all of them (K = total number of stations). The average deviation expresses the mean value of the residuals through the day. A non-zero value is an indication of a systematic bias. Because large positive residuals can cancel with large negative values, a low value of average deviation does not always guarantee close agreement between prediction and observation. By computing the average absolute deviation, however, one can assess whether such a "cancelation" problem is occuring. A large value is an indication of appreciable concentration differences, providing such information even if the average deviation is small. Since a small number of large residuals can dominate in the computation of the previous measures, a large value for either of them does not necessarily indicate consistently large disagreement between prediction and observation. Residuals can be normalized to balance the effect of large and small residuals. This average relative absolute deviation is a measure in whose computation this is done.

The standard deviation, as expressed in Eq. (C-4), is a measure of the shape of the frequency distribution of the residuals. A large value indicates that residual values vary throughout a large range. Correspondingly, a small value suggests that they cluster closely about their mean value, as expressed in Eq. (C-1).

Another statistical measure is of interest. The correlation coefficient, as expressed below, provides an indication of the extent to which variations in observed station concentrations are matched by variations in the predicted station values. A close match is indicated by a value near to one (the value for "perfect" correlation).

- Correlation Coefficient

$$r^m = \frac{\frac{1}{KN-1} \sum_{n=1}^{N} \left[\sum_{k=1}^{K} \left(\tilde{C}_k^{m,n} - \mu_{\tilde{c}}^m \right) \sum_{k=1}^{K} \left(C_k^{m,n} - \mu_c^m \right) \right]}{\sigma_{\tilde{c}}^m \, \sigma_c^m} \qquad \text{(C-6)}$$

where

$$\mu_{\tilde{c}}^m = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} \tilde{C}_k^{m,n}}{KN} \qquad \text{(C-7)}$$

$$\mu_c^m = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} C_k^{m,n}}{KN} \qquad \text{(C-8)}$$

$$\left(\sigma_{\tilde{c}}^m \right)^2 = \frac{1}{KN-1} \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\tilde{C}_k^{m,n} - \mu_{\tilde{c}}^m \right)^2 \qquad \text{(C-9)}$$

$$\left(\sigma_c^m \right)^2 = \frac{1}{KN-1} \sum_{n=1}^{N} \sum_{k=1}^{K} \left(C_k^{m,n} - \mu_c^m \right)^2 \qquad \text{(C-10)}$$

If the value of the correlation coefficient is not close to one, this may or may not be an indication that model performance is deficient. For instance, suppose slight errors were embedded in the wind field supplied to the model. Possibly, the only effect of this could be a slight offset between the predicted and the "true" pollutant cloud location. The concentration level and its distribution within the cloud might be well predicted otherwise. However, the correlation coefficients computed at individual stations (K = 1) might not demonstrate agreement between prediction and observation, indicating instead the opposite. Conceivably, this also might be the case even if the correlation coefficient is computed using concentration values averaged for all stations (K = total number of stations).

Another statistical measure is useful in overcoming this difficulty when sampling stations are not too "sparsely" sited. This measure is the offset correlation coefficient and is designed to compare predictions at one station and time against observations at another station and/or time. It is defined as follows:

> Offset Correlation Coefficient

$$R^{m}_{k,j}\,(\Delta n) = \frac{\dfrac{1}{N-1}\displaystyle\sum_{n=1}^{N}\left(\tilde{C}^{m,n}_{k} - \mu^{m}_{\tilde{C}_{k}}\right)\left(C^{m,n+\Delta n}_{j} - \mu^{m}_{C_{j}}\right)}{\sigma^{m}_{\tilde{C}_{k}}\,\sigma^{m}_{C_{j}}} \tag{C-11}$$

where k is the index of the measurement station at which concentrations are predicted, j is the index of the station at which they are measured, and Δn is the time offset between prediction and observation; also

$$\mu_{\tilde{C}_{k}} = \frac{\displaystyle\sum_{n=1}^{N}\tilde{C}^{m,n}_{k}}{N} \tag{C-12}$$

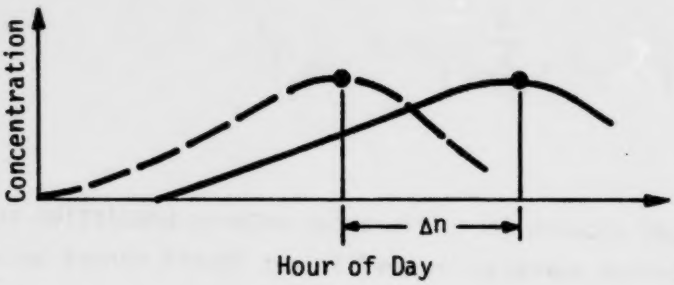
$$\mu_{C_{j}} = \frac{\displaystyle\sum_{n=1}^{N}C^{m,n+\Delta n}_{j}}{N} \tag{C-13}$$

$$\left(\sigma^{m}_{\tilde{c}_k}\right)^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(\tilde{c}_k^{m,n} - \mu_{\tilde{c}_k}\right)^2 \qquad \text{(C-14)}$$

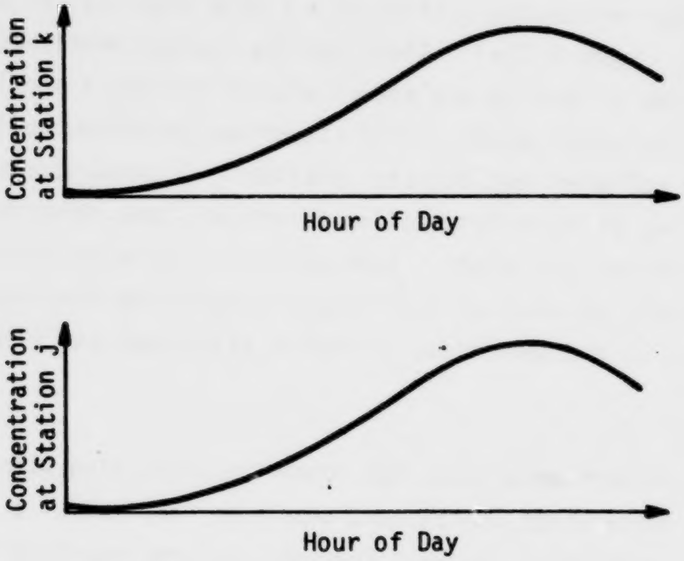
$$\left(\sigma^{m}_{c_j}\right)^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(c_j^{m,n+\Delta n} - \mu_{c_j}\right)^2 \qquad \text{(C-15)}$$

Many reasons can account for differences between prediction and observation. The offset correlation coefficient itself cannot be used to isolate specific reasons, but it can detect time lags or spatial offsets between comparative concentration histories. A time lag might occur because of slight differences between modeled and actual wind speed, diurnal inversion height history, emissions, or atmospheric chemistry, as well as any of a number of other reasons. These differences could manifest themselves at a particular monitoring station as a simple time lag, an example of which is shown in Figure C-2(a). Also, for the reasons mentioned above, as well as differences in modeled and actual wind direction, a spatial offset can occur which could result in the actual and predicted pollutant clouds passing over different but adjacent stations. A comparison of the concentration profiles at these two stations, such as those shown in Figure C-2(b), can reveal the offset. Good agreement could be inferred if the value of the offset correlation coefficient between the concentrations at the two stations, at the same time, assumed a value near one ("perfect" correlation).

In using station data as a basis for comparing prediction with observation, the offset correlation coefficient should be computed as a matter of course. For the station of interest (perhaps the one recording the highest concentration value), computation of the following offset correlation coefficients might be revealing: first, at the same hour, with all adjacent stations (unless none are nearby); then, at the same station, for adjacent hours (for example, one and two hours lag and lead); and finally, with all adjacent stations and hours (to reveal the joint presence of spatial offset and time lag).



(a) Time Lag (Predicted and Measured Concentrations are for the same monitoring station)



(b) Spatial effect (Predicted and Measured Concentrations are for Different but Adjacent Monitoring Stations)

FIGURE C-2. CONCENTRATION HISTORIES REVEALING TIME LAG OR SPATIAL OFFSET

For all the monitoring stations considered together, several other statistics are of interest. For instance, the variation of bias in model predictions with the level of pollutant concentration can be plotted as shown in Figure C-3 . In this particular example, based upon simulations of the Denver Metropolitan region performed using the SAI Urban Airshed Model, the fractional mean deviation from perfect agreement between prediction and observation appears to vary randomly at the higher ozone concentrations. Aside from an apparent systematic bias at very low concentrations, no conclusion of significant bias seems demonstrable.

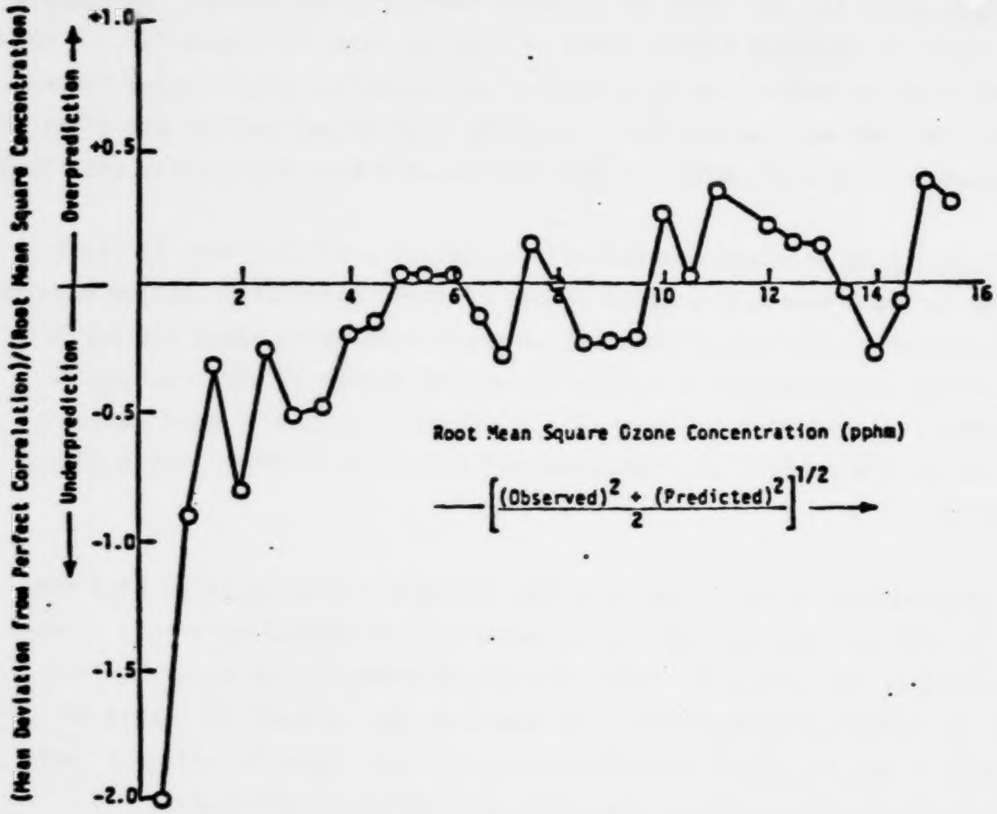


FIGURE C-3. ESTIMATE OF BIAS IN MODEL PREDICTIONS AS A FUNCTION OF OZONE CONCENTRATION. This figure is based upon predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.

Residuals can vary in sign and magnitude during the modeling day. It is often helpful to plot their diurnal variation. An example is shown in Figure C-4, based upon predictions of the SAI Urban Airshed Model for three modeling days in Denver. A discernable pattern might be symptomatic of basic model inadequacies. In this example, however, no simple pattern seems apparent.

For each set of observations or predictions (for all stations and times), there exists a cumulative concentration frequency distribution. This describes the probability of occurence of a concentration in excess of a certain value for the range of possible concentration values. An example based upon the modeling effort noted earlier in shown in Figure C-5. A conclusion might be drawn from this figure: Although background ozone concentrations are not well-determined (low background concentrations are difficult to measure accurately), higher concentrations are more predictably distributed.

By plotting observed concentrations against predicted ones (at each station for each hour), a graphic record of their correlation can be obtained. The degree of clustering of observation-prediction pairs about the perfect correlation line provides an indication of the degree of their agreement. An example is presented in Figure C-6. For each particular combination of observation and prediction, the number of occasions on which they occurred are shown.

Superimposed on the figure are the standard deviation bands (1σ) for both the EPA standard and maximum acceptable instrumentation error. These bands portray the extent to which station measurements are accurate indicators of "true" concentrations. To conclude that a model is unable to reproduce a set of "true" concentrations, one must know the value of those concentrations. Measurements, however, are imperfect surrogates. If concentration residuals are within instrumentation limits, differences could be explained solely by measurement errors. In such a case, no further conclusions could be reached about model predictive ability.

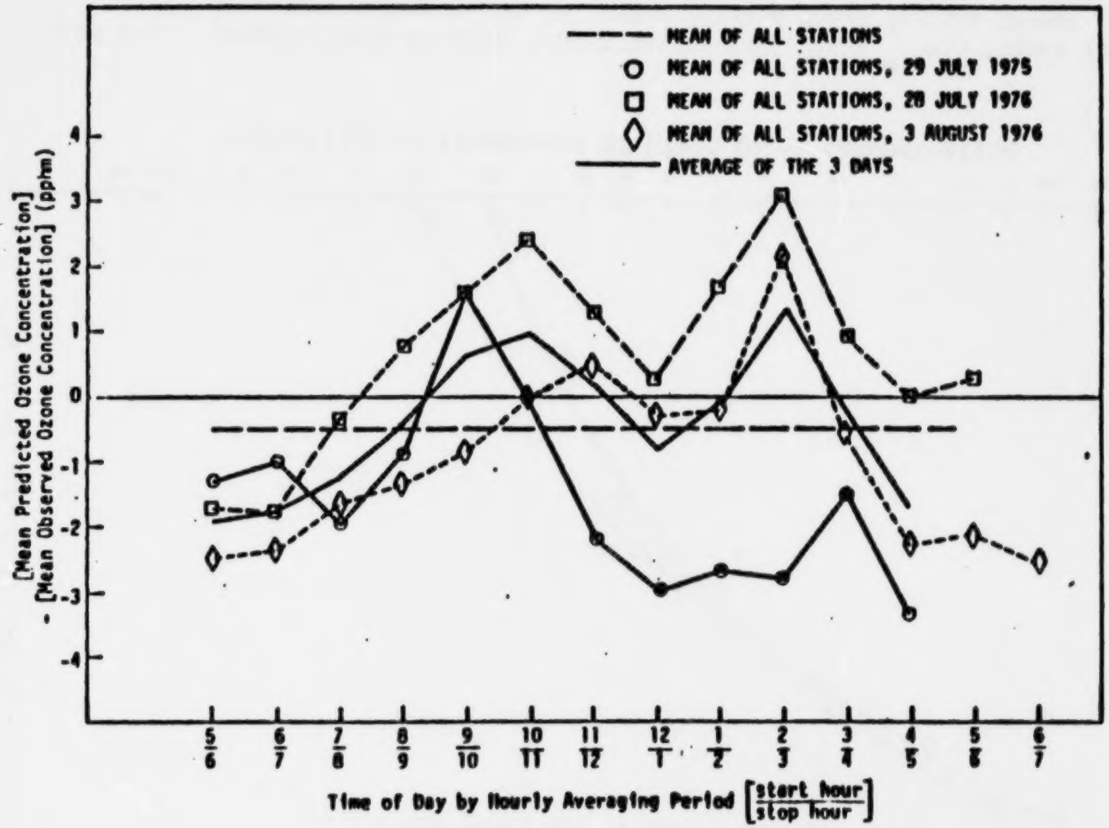


FIGURE C-4. TIME VARIATION OF DIFFERENCES BETWEEN MEANS OF OBSERVED AND PREDICTED OZONE CONCENTRATIONS. This figure is based upon predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.

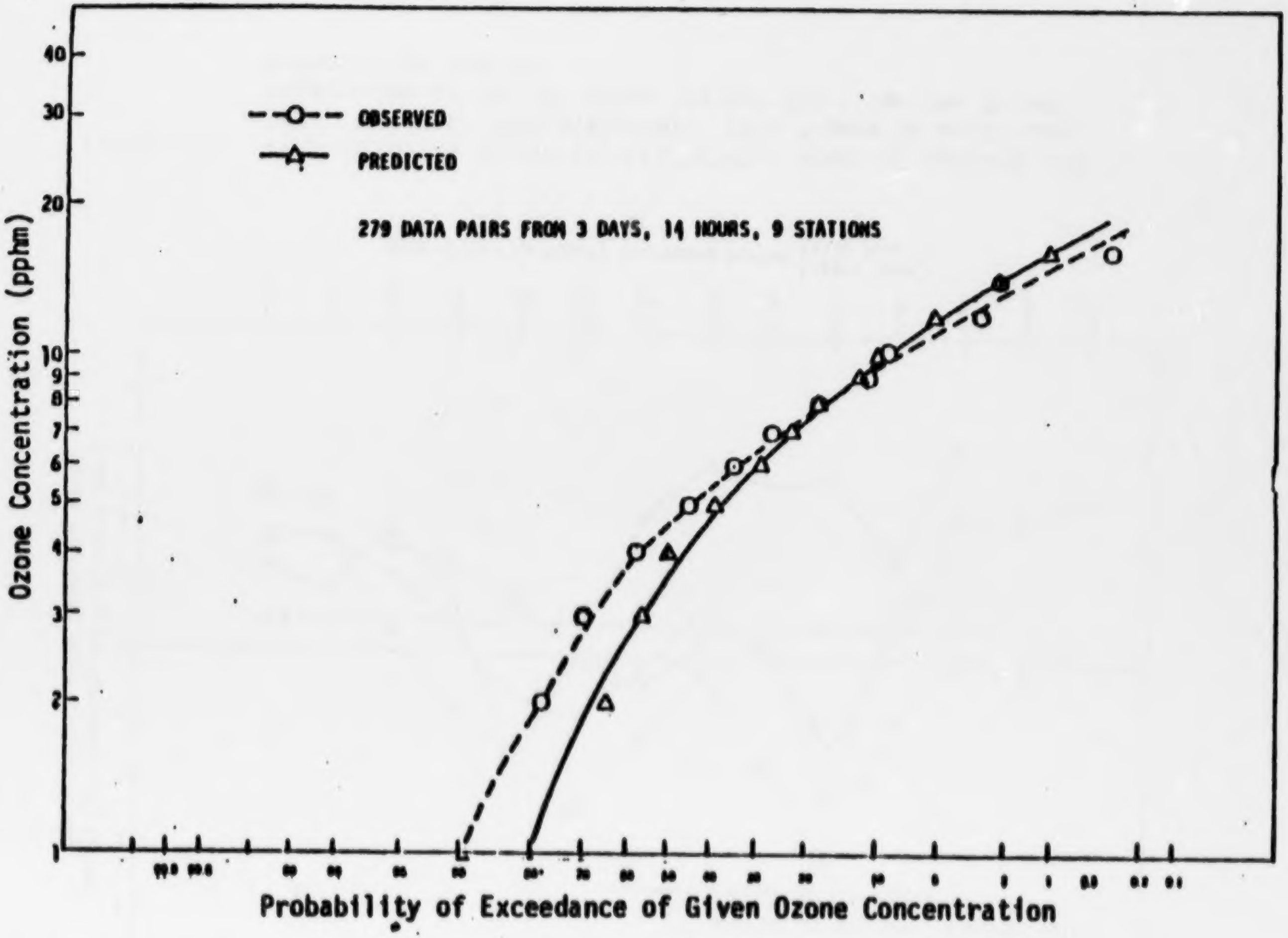


FIGURE C-5. PROBABILITIES OF OZONE CONCENTRATION EXCEEDANCE. This figure is based upon predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.

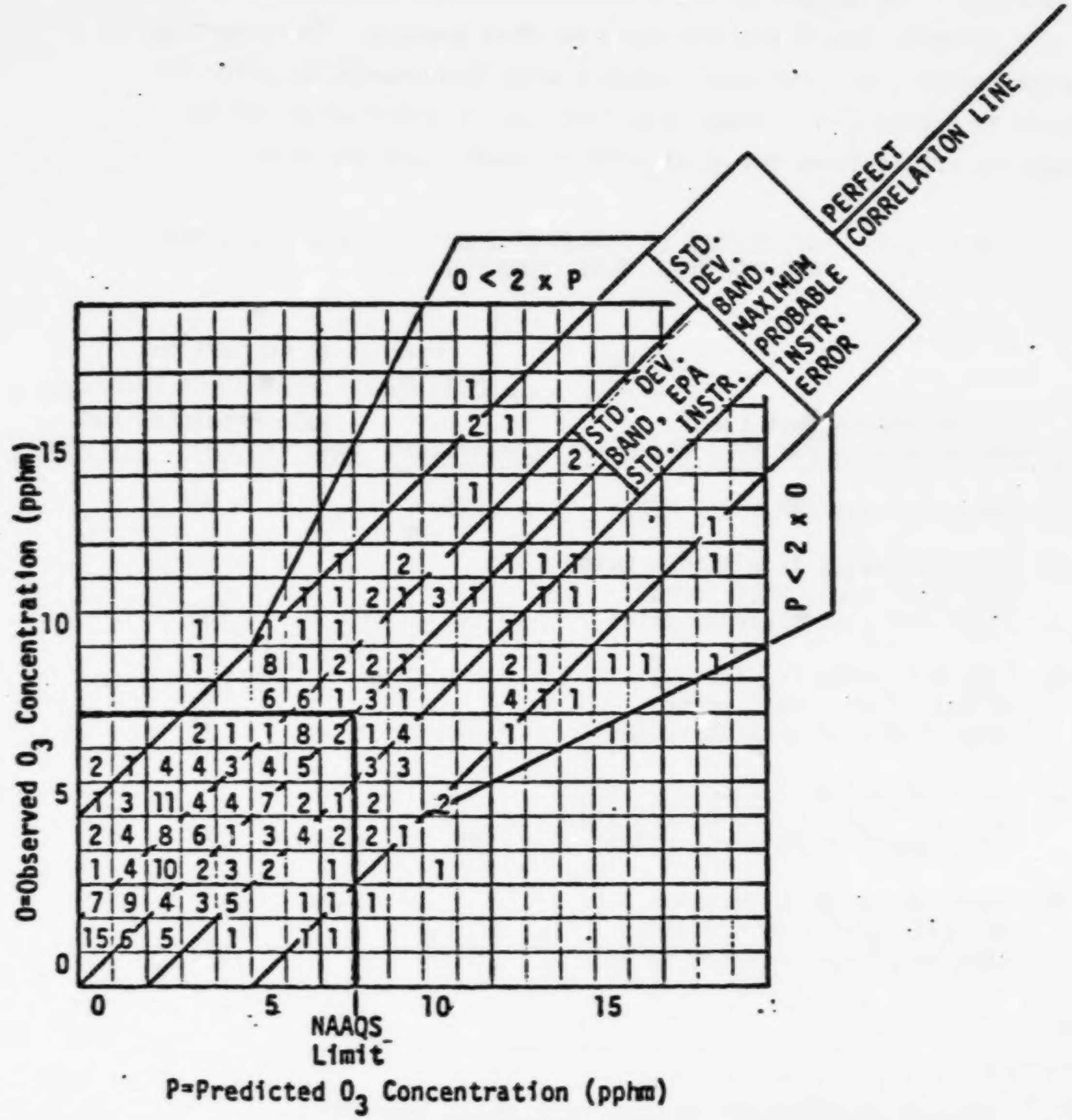


FIGURE C-C. MODEL PREDICTIONS CORRELATED WITH INSTRUMENT OBSERVATIONS OF OZONE (DATA FOR 3 DAYS, 9 STATIONS, DAYLIGHT HOURS). This figure is based on predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.

Some of the information contained in Figure C-6 is summarized in Table C-4. The percent of prediction/observation pairs meeting certain correspondence levels are indicated for this example. The extent to which concentration residuals compare with instrumentation error is shown in Figure C-7. These same plots can be constructed for most modeling applications for which station predictions are known.

TABLE C-4. OCCURRENCE OF CORRESPONDENCE LEVELS OF PREDICTED AND OBSERVED OZONE CONCENTRATIONS

Correspondence Level Between Predicted and Observed Pairs	Percent of Comparisons Meeting Correspondence Level	
	Comparisons	Both Predicted and Observed Conc. > 8 pphm
1) Factor of two (2P > O > P/2)	80%	94%
2) Computed value is within ± twice S.D. max. prob. inst. error (95% level) of observed value	100	100
3) Computed value is within ± S.D. of max. prob. inst. error (95% level) of observed value	93	90
4) Computed value is within ± twice S.D. of inst. errors by EPA std. (95% level) of observed value	89	77
5) Computed value is within ± S.D. of inst. errors by EPA std. (95% level) of observed value	60	37

c. "Pattern Recognition" Station Performance Measures

Several qualitative/composite model performance measures are useful in comparing station predictions with observations. At each monitoring site, for instance, the time history through the modeling day of the predicted concentrations can be plotted directly with the time history of

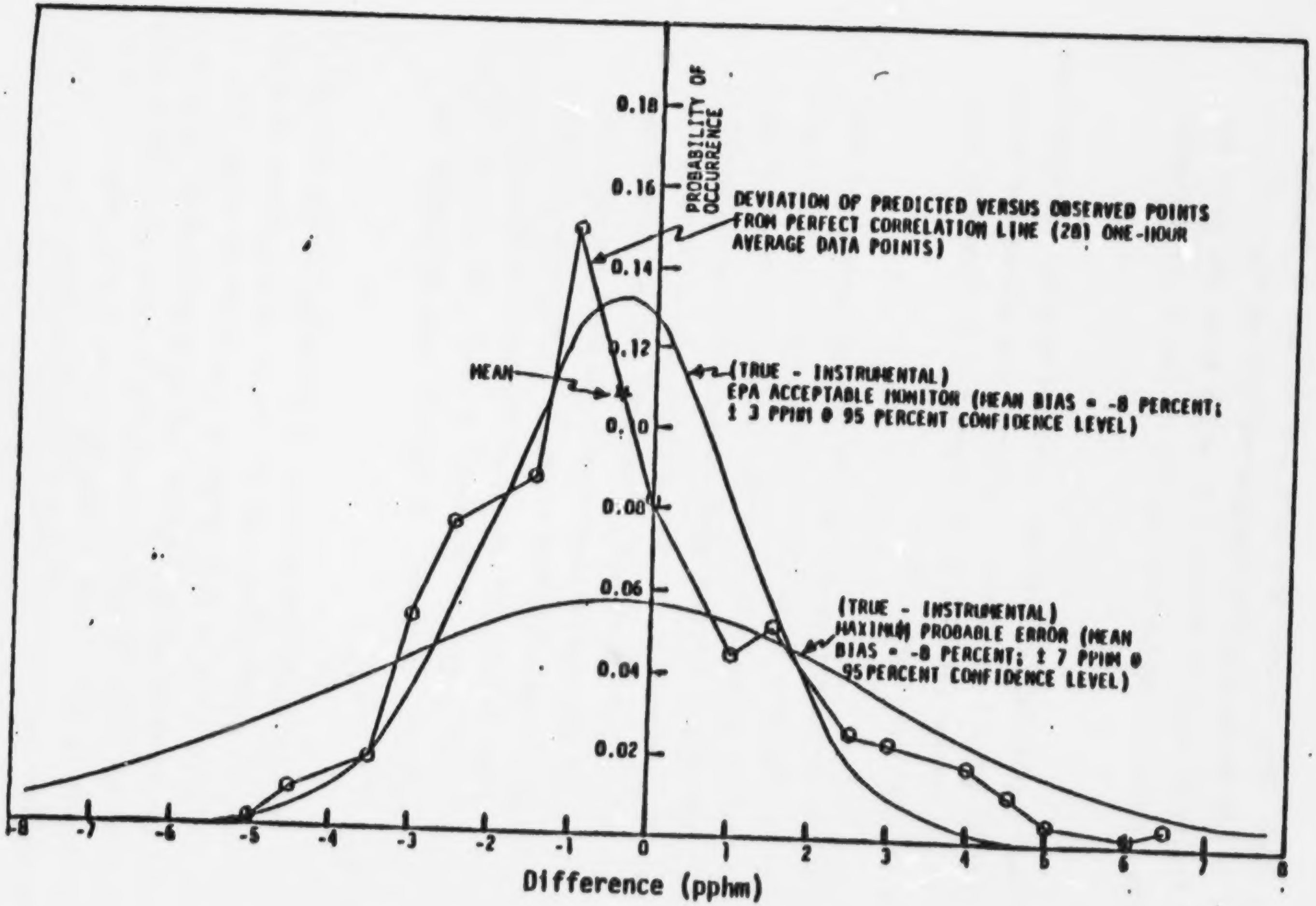


FIGURE C-7. MODEL PREDICTIONS COMPARED WITH ESTIMATES OF INSTRUMENT ERRORS FOR OZONE (DATA FOR 3 DAYS, 9 STATIONS, DAYLIGHT HOURS) the measurement data. This is done in Figure C-9 for one of the days (3 August 1976) in the Denver modeling example employed earlier. Preceding this figure is a map in Figure C-8, which shows the names and locations of the air quality monitoring stations in the Denver Metropolitan region.

For each hour during the day, the predicted and observed concentrations each can be averaged for all measurement stations. The diurnal variation of this all-station average can also be of interest. An example of such a time history is shown in Figure C-10.

At the time the concentration peak occurs, the performance of the model in predicting that peak is of interest as is its ability to predict the lower concentration values at monitoring stations distant from the peak. An indication of the relative prediction-observation agreement at the peak versus the agreement at outlying stations can be found by computing a composite performance measure. The ratio can be found of the normalized residual at the station measuring the highest concentration value to the average of the normalized residuals at the other stations. If this ratio is large, better performance at the outlying stations than near the peak can be inferred. If the value is small, the reverse is true. If the ratio is near unity, agreement is much the same throughout the modeled region.

The value of a concentration residual at a station changes during the modeling day. If these changes can be tied to corresponding changes in atmospheric characteristics (the height of the inversion base, for instance), we can sometimes draw valuable inference about model performance as a function of the value of these atmospheric "forcing variables." Some of these variables include: wind speed, inversion height, ventilation (combining the previous two variables into a product of their values), solar insolation, and a particular category of emissions (automotive, for example).

KEY

NG - Northglenn
WE - Welby
AR - Arvada
CR - C.A.R.I.H.
CM - Continuous Air Monitoring Program [CAMP]
NJ - National Jewish Hospital
GM - Green Mountain
OV - Overland
PR - Parker Road

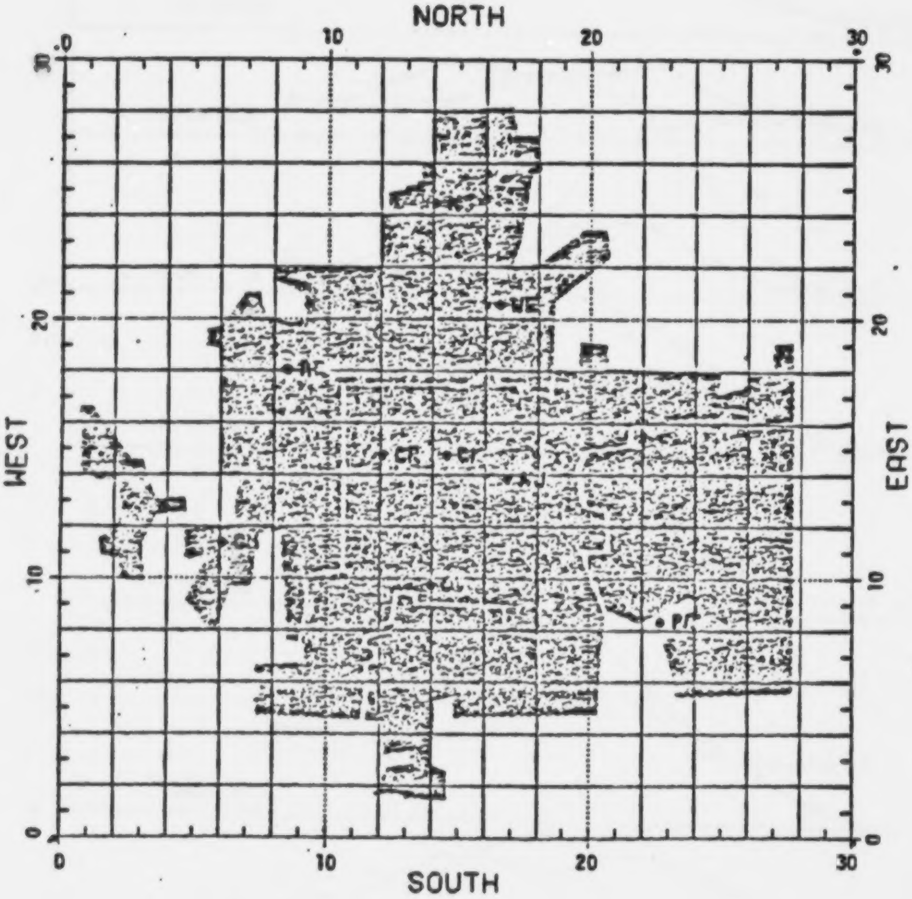

NORTH
SOUTH
WEST
EAST
0
10
20
30
WE
CR
PR

FIGURE C-8. MAP OF DENVER AIR QUALITY MODELING REGION SHOWING AIR QUALITY MONITORING STATIONS

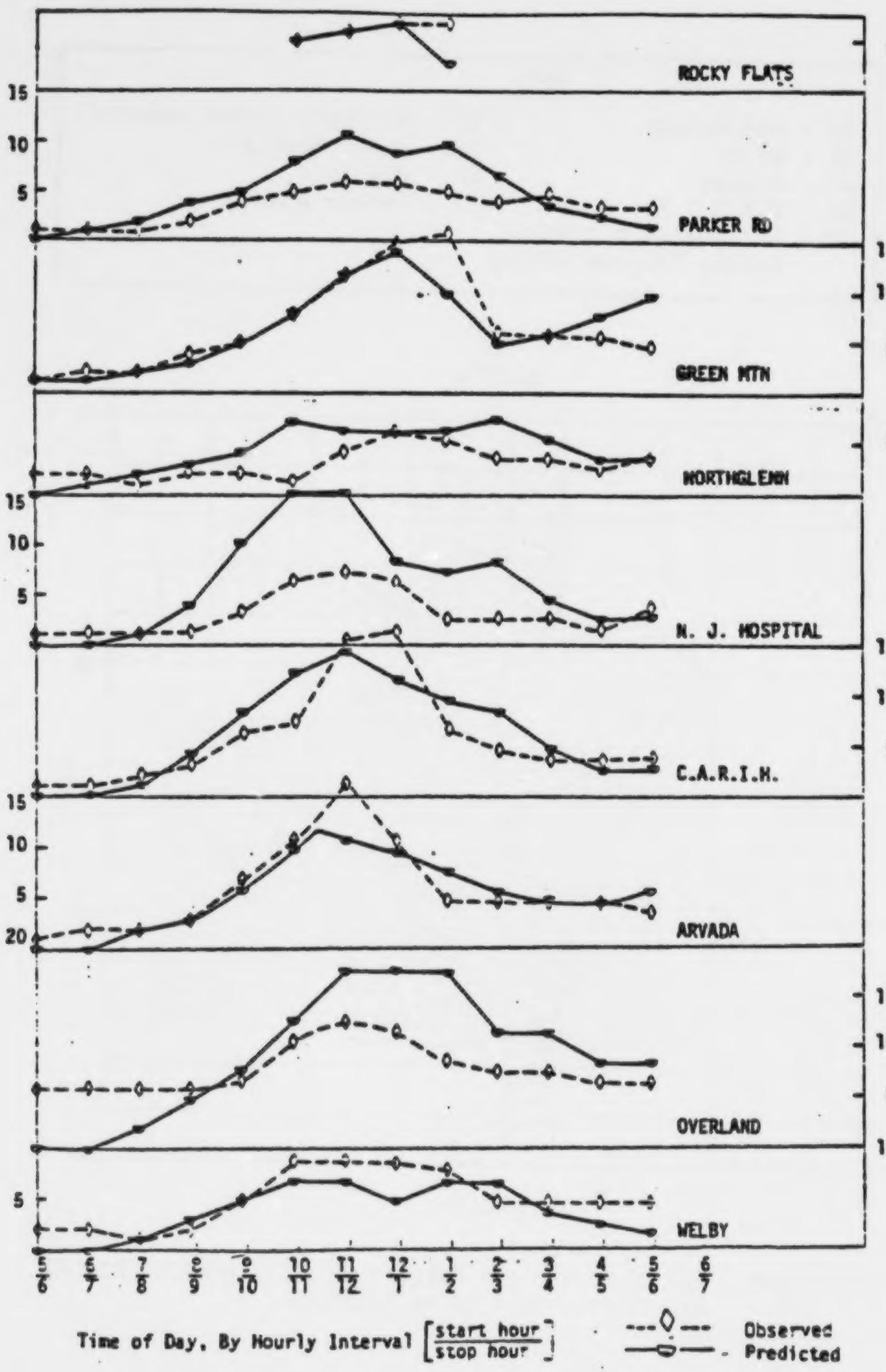


FIGURE C-9. TIME HISTORY OF PREDICTED AND OBSERVED CONCENTRATIONS AT MONITORING SITES. This figure is based on the predictions of the SAI Urban Airshed Model in Denver for 3 August 1976.

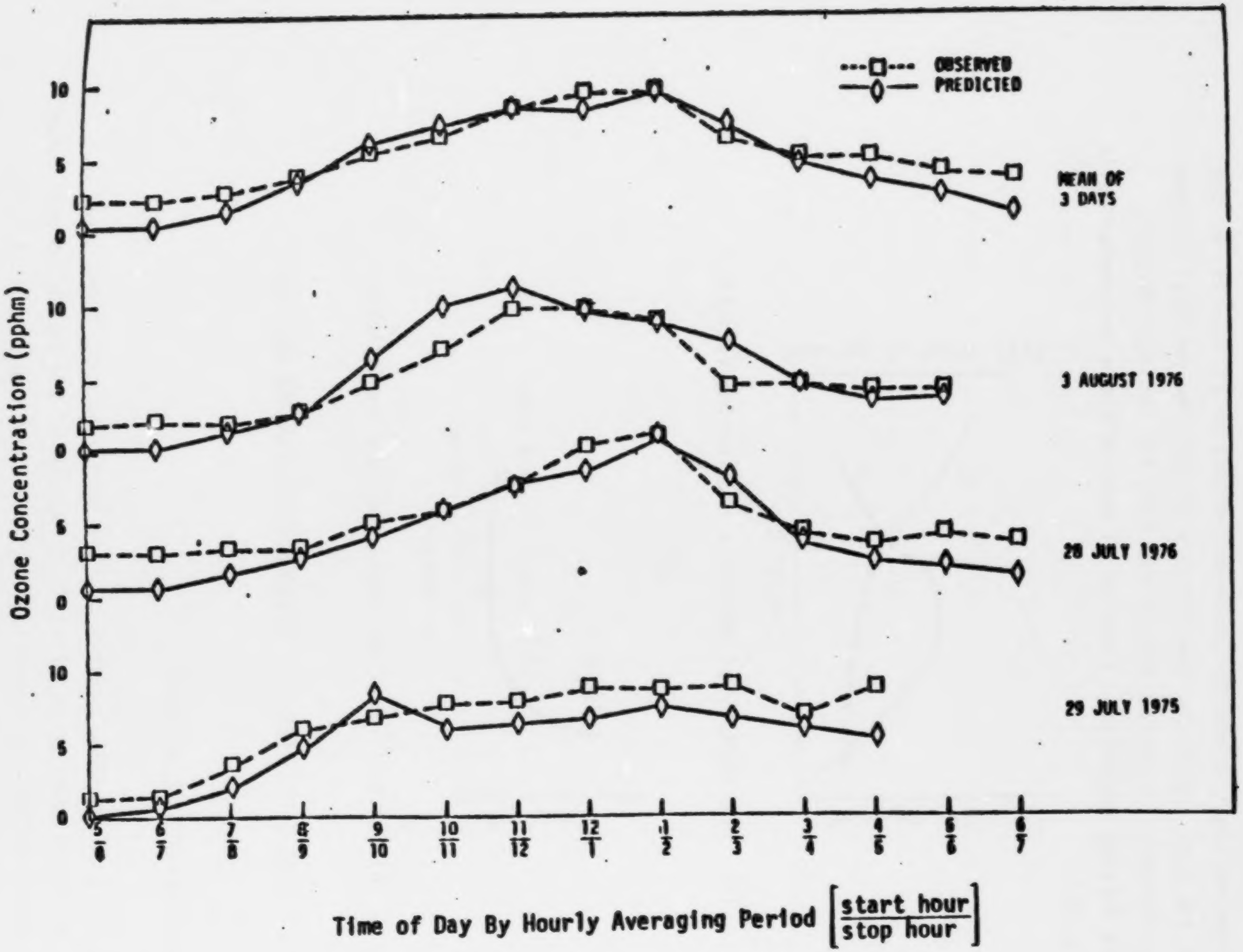
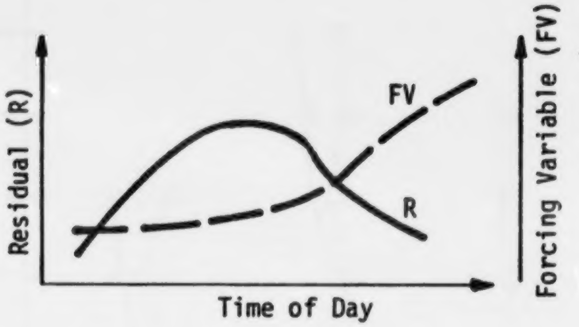
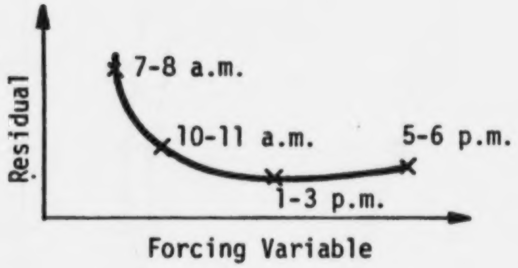


FIGURE C-10. VARIATIONS OVER ALL STATIONS OF OBSERVED AND PREDICTED AVERAGE OZONE CONCENTRATIONS
This figure is based on the prediction of the SAI Urban Airshed Model in Denver.

To examine residual values for cause-and-effect relationships, we can plot on the same figure the time history of both the residual and the forcing variable. Alternatively we can plot the residual directly with the forcing variables. Examples of both of these are presented in Figure C-11.



(a) Time History of Residuals and Forcing Variable



(b) Cross-Plot of Residuals and Forcing Variable

FIGURE C-11. PLOTS OF RESIDUALS AND FORCING VARIABLE

3. AREA PERFORMANCE MEASURES

To use a performance measure of this type, one must know, usually at each daylight hour, the spatial distribution of the predicted and "true" concentration fields. By comparing the two, either throughout the day or at a specific time or event, we can construct in principle, a number of model performance measures. In practice, however, we are seldom able to resolve fully the "true" concentration field, even if the model we use is capable of doing so for the predicted field. This difficulty derives from the limited sampling of measurement data generally available: Only measurements at several scattered monitoring stations are recorded. Unless ambient conditions are highly predictable and the monitoring network is extensive and exceptionally well-designed, reconstruction of the "observed" concentration field from discrete station measurements can be an uncertain and error prone process.

Nevertheless, the observed concentration field can be inferred with accuracy in some cirucmstances. In addition, models frequently can provide spatially resolved predictions. Grid models, for instance, predict average concentrations in a number of grid cells. Resolution is then provided as finely as the horizontal grid-cell dimensions (on the order of one to several kilometers). Trajectory model predictions can be used to calculate concentrations along the space-time track followed by the air parcel being modeled. Gaussian models are analytic and can resolve fully their predictions. Thus, even if the observed concentration field is known only imperfectly, the predicted field, because it is often much better resolved, can still provide qualitative information about model performance. Further, the shape of the predicted concentration field can suggest ways to extract information for comparison with station measurements. We discuss "hybrid" performance measures later in this Appendix.

In this section we present several area performance measures. When predicted and observed concentration fields are known, they can provide considerable insight into model performance. These performance measures are based upon taking the difference between the predicted and observed values of certain quantities. Even when the observed values of these quantities are not known with accuracy, computation of their predicted values can provide a systematic means for characterizing model predictions.

The performance measures presented here can be divided into three types: scalar, statistical, and "pattern recognition." We discuss each in turn. In Table C-5, we list some of these measures.

a. Scalar Area Performance Measures

The seriousness of a pollutant problem is a function not only of the concentration level itself but also of the spatial extent of the pollutant cloud. Several scalar area performance measures are designed with this in mind. Even if a model predicts the peak concentration well, it may not necessarily predict the extent of the area exposed to concentrations near to that value. This might not be a serious defect if the pollutant cloud passed over uninhabited terrain. However, if the cloud were to drift over a densely populated urban area, a considerable difference in the health effects experienced could exist between a cloud one mile across and another five miles across. This could affect correspondingly our willingness to accept a model for use whose predictions of cloud dimensions differed considerably from observed dimensions.

Two performance measures of interest are the following: the differences between both the fraction of the area of interest within which concentrations exceed the NAAQS and the fraction experiencing concentrations within 10 percent of the peak value. The first of these is a measure of the general ability of the model to predict the spatial extent of concentrations in the range of interest. The second estimates the performance of the model in the higher concentration ranges at which, presumably, health effects are more pronounced.

A third measure is of interest. At each measurement station a set of concentration readings are recorded. It is interesting to compute from the predicted concentration field the nearest distance at which there occurs a value equal to the observed value, as well as the azimuthal direction from the station to the nearest such point. This direction lies along the concentration gradient of the predicted field. The magnitude of the distance is a measure of the spatial offset between the predicted and observed concentration fields in the vicinity of the monitoring station. The direction is a measure of the orientation of the offset.

TABLE C-5. SOME AREA PERFORMANCE MEASURES

Type	Performance Measure
Scalar	a. Difference in the fraction of the area of interest in which the NAAQS are exceeded.
	b. Nearest distance at which the observed concentration is predicted.
	c. Difference in the fraction of the area of interest in which concentrations are within 10 percent of the peak value.
Statistical	a. At the time of the peak, differences in the fraction of the area experiencing greater than a certain concentration; differences in the following are of interest: 1) Cumulative distribution function 2) Density function 3) Expected value of concentration 4) Standard deviation of density function
	b. For the entire residual field, the following statistics are of interest: 1) Average deviation 2) Average absolute deviation 3) Average relative absolute deviation 4) Standard deviation 5) Correlation coefficient 6) Estimate of bias as a function of concentration 7) Comparison of the probabilities of concentration exceedances as a function of concentration
	c. Scatter plots of prediction-observation concentration pairs with a line of best fit determined in a least squares sense.
Pattern recognition	a. Isopleth plots showing lines of constant pollutant concentration for each hour during the modeling day.
	b. Time history of the size of the area in which concentrations exceed a certain value.
	c. Isopleth plots showing lines of constant residual values for each hour during the day ("subtract" prediction and observed isopleths).
	d. Isopleth plots showing lines of constant residuals normalized to selected forcing variables (inversion height, for instance).
	e. Peak-to-overall performance indicator, computed by taking the ratio of the mean residual in the area of the peak (e.g., where concentrations are within 10 percent of the peak) to the mean residual in the overall region.

b. Statistical Area Performance Measures

A number of statistical area performance measures are of use. They are generally computed either at a fixed time or at the time of a fixed event, (the peak,for instance). Before they can be computed, however, both the predicted and observed concentration field must be transformed into a compatible, discrete form. The scales of resolution must be made the same, though kept as fine as possible. For example, if a grid model provided average concentrations every two kilometers in a lattice-work pattern spanning the region of interest, then the observed concentration field inferred from station measurement must also be resolved at two kilometer intervals with concentrations obtained at each point in the lattice-work. If resolution cannot be obtained so finely, then the predicted concentration field must be adjusted to be comparable with the observed one. The field having the coarsest resolution if the limiting one.

Once the fields have been resolved into a compatable form, several performance measures can be computed. We can characterize a concentration field by indicating for each concentration value the fraction of the area experiencing a concentration greater than that value. By so doing, we define a cumulative distribution function (CDF) such as that shown in Figure C-12. The CDF is the integral of its density function (f), also shown in the figure.

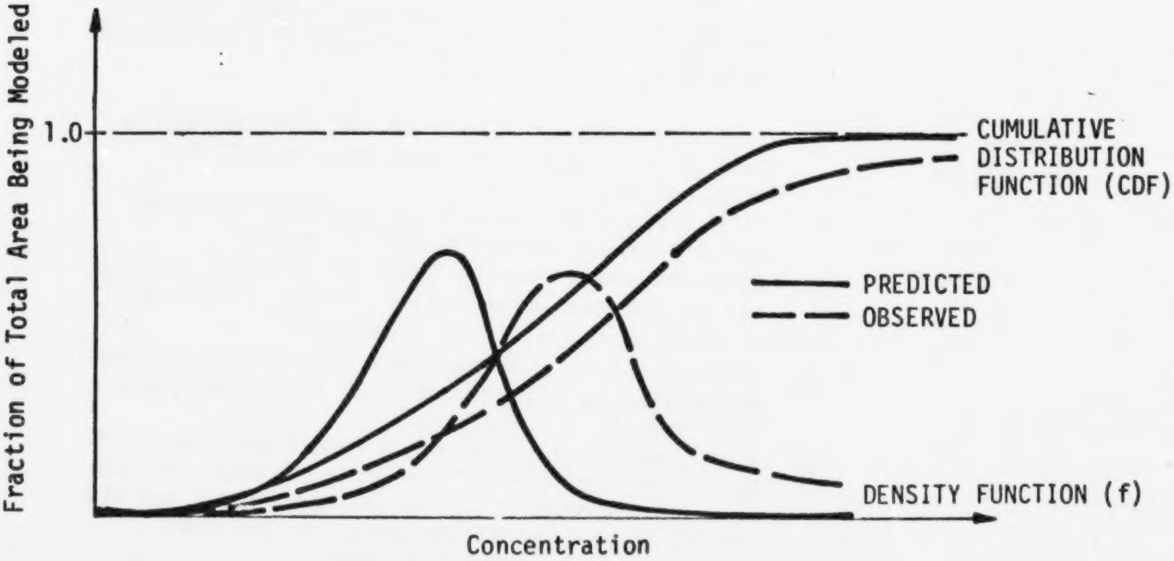


FIGURE C-12. DISTRIBUTION OF AREA FRACTION EXPOSED TO GREATER THAN A GIVEN CONCENTRATION VALUE

For the predicted and observed concentration fields, the CDF's may differ. The following statistics can be compared in order to characterize the difference: the CDF itself, the mean expected concentration in the modeled region, and the standard deviation of the area density function. If the CDF and f were continuous functions, the following express the form of these measures:

> Cumulative Distribution Function

$$CDF(C \leq K) = \int_{C_B}^{K} f(c)dc \qquad (C\text{-}16)$$

> Expected Concentration

$$\mu_A = \int_{C_B}^{C_P} cf(c)dc \qquad (C\text{-}17)$$

where C_P is the peak and C_B is the background concentration.

> Standard Deviation

$$\sigma_A^2 = \int_{C_B}^{C_P} (c - \mu_A)^2 f(c)dc \qquad (C\text{-}18)$$

However, the CDF and f are not available in practice as continuous functions: They are expressed discretely, derived from concentrations at the nodal points of a ground-level grid having dimensions I by J. The above measures have the following discrete form:

> Discrete Cumulative Distribution Function

$$CDF(C^m \leq K) = \frac{1}{IJ} \sum_{j=1}^{J} \sum_{i=1}^{I} u(K - C_{ij}^m) \qquad (C\text{-}19)$$

where m is the pollutant species and u is a unit step function whose value is

$$u(x) = \begin{cases} 1 , & x \geq 0 \\ 0 , & x < 0 \end{cases} \tag{C-20}$$

> Discrete Expected Concentration

$$\mu_A^m = \frac{1}{IJ} \sum_{j=1}^{J} \sum_{i=1}^{I} C_{ij}^m \tag{C-21}$$

> Discrete Standard Deviation

$$\left(\sigma_A^m\right)^2 = \frac{1}{IJ - 1} \sum_{j=1}^{J} \sum_{i=1}^{I} \left(C_{ij}^m - \mu_A^m\right)^2 \tag{C-22}$$

The predicted and observed concentration fields can be differenced, with the result being a spatially distributed residual field at the fixed time or event of interest. The statistics of this residual field are essentially the same as those described earlier in Eqs. (C-1) to (C-10) for the set of station residuals. They are as follows (the tilde ˜ denotes "predicted," while m is the pollutant species and I, J are the number of nodes in the concentration field grid):

> Average Deviation

$$\mu^m = \frac{1}{IJ} \sum_{j=1}^{J} \sum_{i=1}^{I} \left(\tilde{C}_{ij}^m - C_{ij}^m\right) \tag{C-23}$$

> Average Absolute Deviation

$$|\mu|^m = \frac{1}{IJ} \sum_{j=1}^{J} \sum_{i=1}^{I} |\tilde{C}_{ij}^m - C_{ij}^m| \tag{C-24}$$

Average Relative Absolute Deviation

$$|\bar{\mu}|^{m} = \frac{1}{IJ}\sum_{j=1}^{J}\sum_{i=1}^{I}\frac{|\tilde{C}_{ij}^{m} - C_{ij}^{m}|}{C_{ij}^{m}} \qquad \text{(C-25)}$$

Standard Deviation

$$(\sigma^{m})^{2} = \frac{1}{IJ - 1}\sum_{j=1}^{J}\sum_{i=1}^{I}\left[\left(\tilde{C}_{ij}^{m} - C_{ij}^{m}\right) - \mu^{m}\right]^{2} \qquad \text{(C-26)}$$

Correlation Coefficient

$$r^{m} = \frac{\frac{1}{IJ - 1}\left(\sum_{j=1}^{J}\sum_{i=1}^{I}\tilde{C}_{ij}^{m} - \mu_{\tilde{C}}^{m}\right)\left(\sum_{j=1}^{J}\sum_{i=1}^{I}C_{ij}^{m} - \mu_{C}^{m}\right)}{\sigma_{\tilde{C}}^{m}\,\sigma_{C}^{m}} \qquad \text{(C-27)}$$

Calculation of the above statistics can be extended through the modeling day by including residual values not just at a specific time or event but for each hour during the day. Also, a graphical representation of the correlation between prediction and observation can be developed by plotting prediction-observation concentration pairs on a scatter plot, much as was done for station values in Figure C-6.

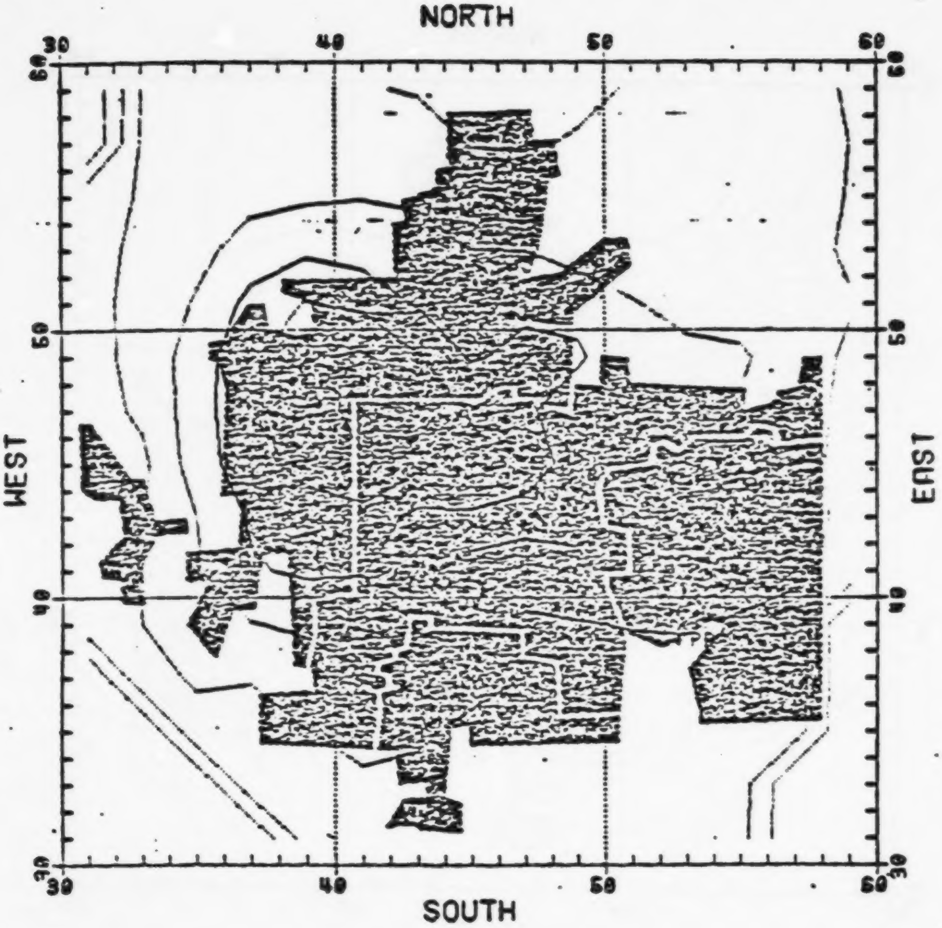
c. "Pattern Recognition" Area Performance Measures

Considerable information about model performance often can be found through the use of "pattern recognition" area performance measures. Even if a comparison between prediction and observation is difficult due to the sparsity of the latter data, insight can still be gained through the use of the measures described here.

The spatial and temporal development of the pollutant cloud is of considerable interest. Frequently, differences between prediction and observation can be spotted quickly by comparing isopleth plots showing contours of constant pollutant concentrations. The development of the cloud can be portrayed graphically in a series of hourly isopleth plots. Shown in Figures C-13(a) through (e) is a series of hourly isopleth plots. These represent predictions for ozone generated by the SAI Urban Airshed Model for the Denver Metropolitan region on 29 July 1975. The locations of the measurement stations are also shown, as they were in Figure C-8.

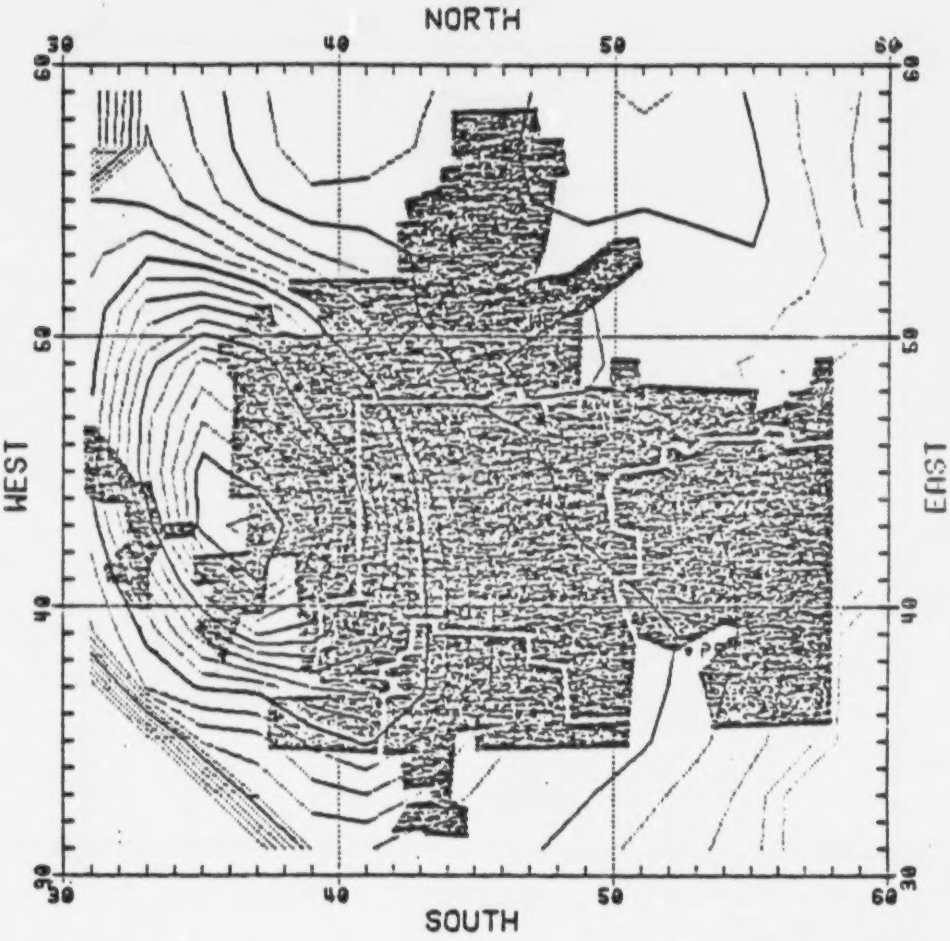
The example illustrated in Figure C-13 is typical of applications involving multiple-source, region-oriented issues (SIP/C, AQMP). However for specific-source issues, the downwind isopleth contours are approximately elliptical. An example of a specific-source isopleth, or "footprint", plot was presented earlier in Figure V-4 in Chapter V.

Model performance can also be characterized by comparing against observation the time histories of the size of the area in which concentrations exceed a certain value. Such a comparison would provide insight into the temporal variation of prediction-observation differences. An example of such a history is presented in Figure C-14 for ozone in the Denver Metropolitan region. A meteorology the same as that observed on 28 July 1976 was employed by the SAI Urban Airshed Model, along with emissions for that date and projected emissions for 1985 and 2000, to predict the spatial and temporal distribution of ozone for each year. Lines of constant concentration values are also shown.


NORTH
SOUTH
WEST
EAST
30
40
50
60

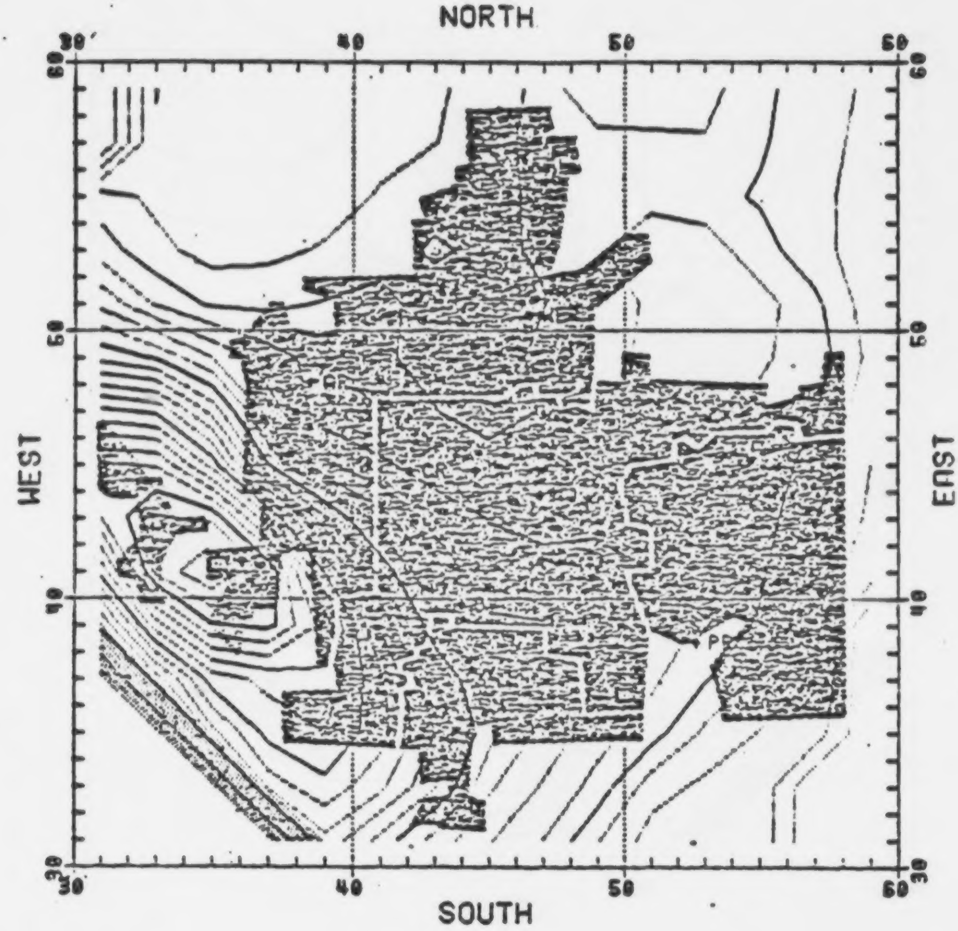
(a) Hour 0800-0900 MST

FIGURE C-13. ISOPLETHS OF OZONE CONCENTRATIONS (pphm) ON 29 JULY 1975. Isopleth interval 1 pphm. This figure is based on predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.


NORTH
SOUTH
WEST
EAST
30
40
50
60
30
40
50
60
30
40
50
60
30
40
50
60

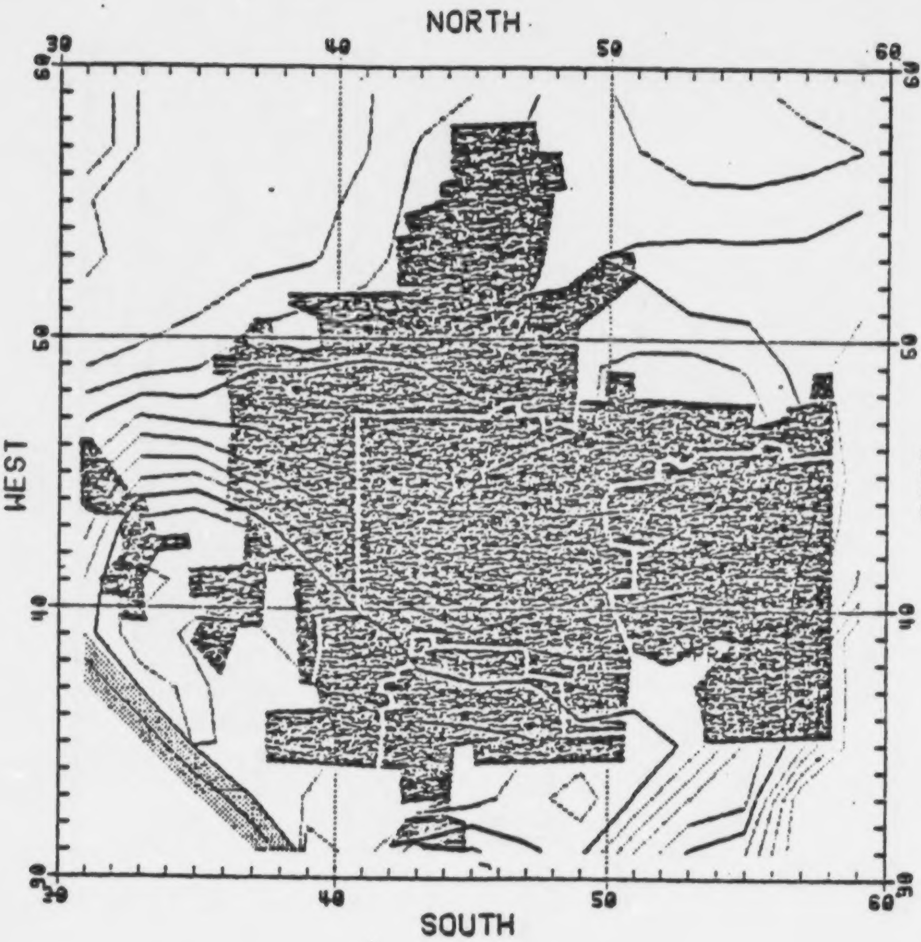
(b) Hour 1000-1100 MST

FIGURE C-13 (Continued)


NORTH
SOUTH
WEST
EAST
30
40
50
60
30
40
50
60
PP

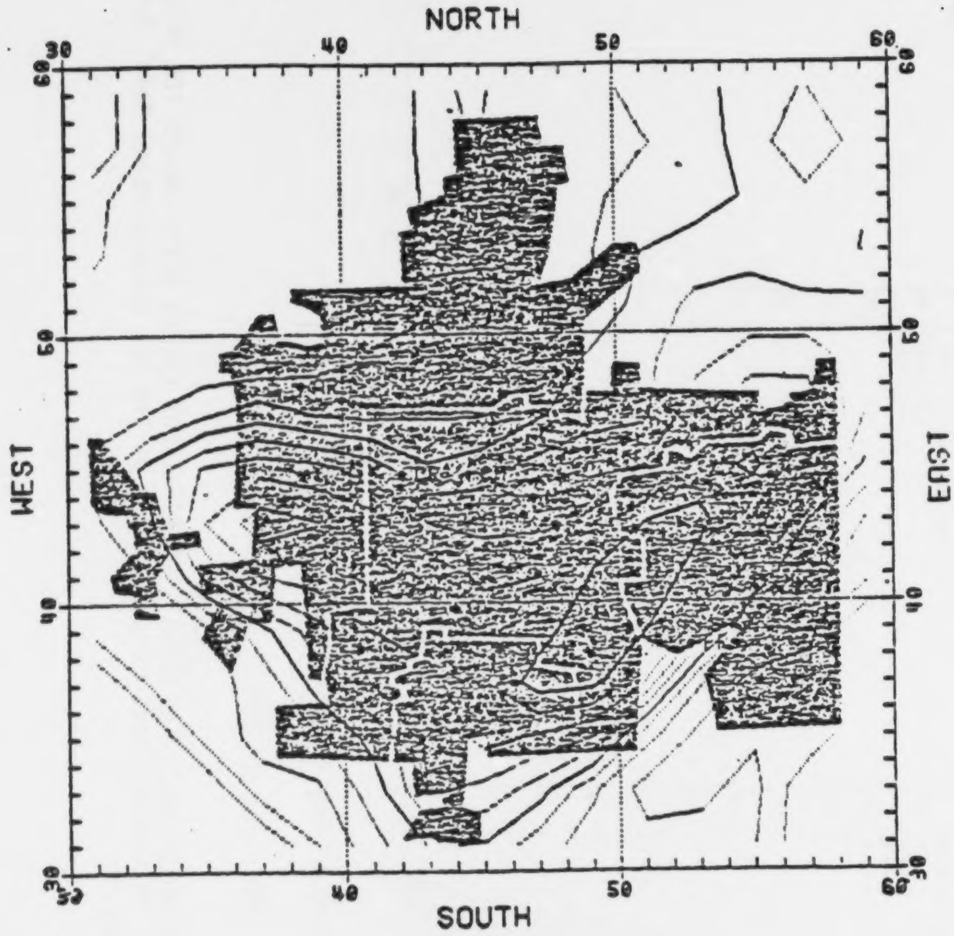
(c) Hour 1200-1300 MST

FIGURE C-13 (Continued)


NORTH
SOUTH
WEST
EAST
30
40
50
60

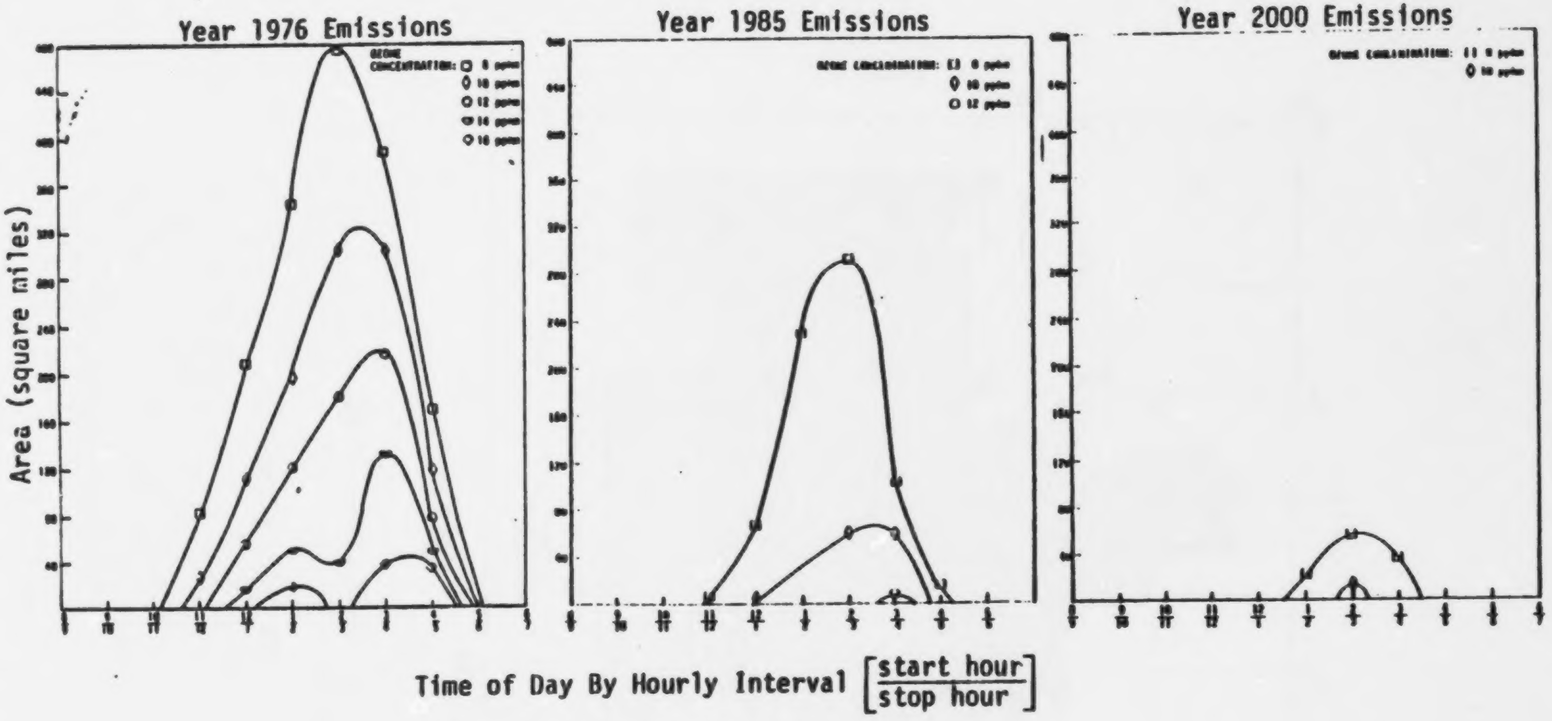
(d) Hour 1400-1500 MST

FIGURE C-13 (Continued)


NORTH
SOUTH
WEST
EAST
30
40
50
60
30
40
50
60
30
40
50
60
30
40
50
60

(e) Hour 1600-1700 MST

FIGURE C-13 (Concluded)



Meteorology for 28 July 1976 Assumed

FIGURE C-14. SIZE OF AREA IN WHICH PREDICTED OZONE CONCENTRATIONS EXCEED GIVEN VALUES FOR YEARS 1976, 1985, AND 2000. This figure is based on predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.

If both the predicted and observed concentration fields are resolved compatibly to the same scale, the two can be differenced and the residuals plotted directly as isopleth contour plots. This may be done either at a fixed time/event or hourly. The example shown in Figure C-15 is typical of such a plot, although it was not derived from observational data. This particular figure was calculated by differencing the annual NO_2 concentrations predicted by the EPA's Climatological Dispersion Model (CDM) for two emissions regions: one a base case and the other a 17.5 percent reduction in emissions in downtown Denver. Since the magnitude of the residuals may be strongly a function of certain atmospheric forcing variables (wind speed or inversion height, for instance), it can be helpful to normalize residuals to the forcing variable values.

Several model performance problems can be spotted qualitatively using residual isopleth plots. Some of those that might be apparent are:

- Good peak/poor spatial agreement.
- Bad peak/good spatial agreement.
- Different peak location.

A composite measure can also be useful in assessing the relative peak/spatial performance of a model. The peak-to-overall indicator can be calculated at the time of the peak as the ratio of the mean residual in the vicinity of the peak (where concentrations are within 10 percent of the peak, for example) to the mean residual in the overall region.

4. EXPOSURE/DOSAGE PERFORMANCE MEASURES

The health effects experienced by an individual in a pollutant region seem to be a function of both the concentration level and the duration of exposure. The aggregate impact experienced by the total populace would be expressed by the sum of the effects impacting each individual. The seriousness of the pollutant problem would be related not just to the spatial and temporal development of the pollutant alone but also to the spatial and temporal distribution of the population living beneath it. Several performance measures attempt to guage model performance on this basis.

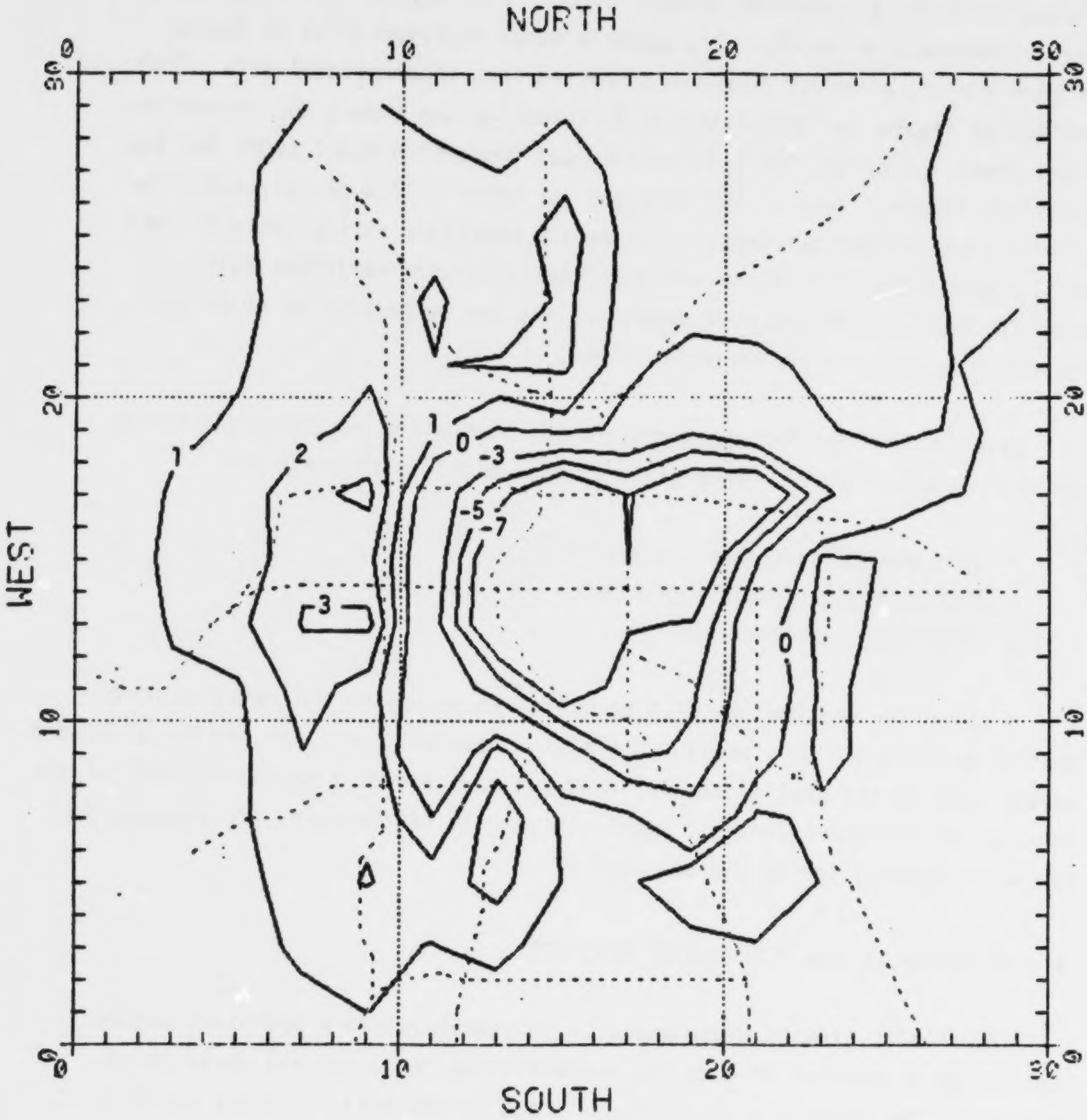


FIGURE C-15. TYPICAL RESIDUALS ISOPLETH PLOT FOR ANNUAL AVERAGE NO_2. Units are in μg/m^3.

In this section we present some of these performance measures, acknowledging at the outset the difficulty of their computation in practice. Whether the spatial scale is urban/regional or source-specific, the problem is essentially the same. Not only must the predicted and observed concentration field be known, but also the population distribution. All are temporally and spatially varying. Conceivably, the observed concentration field may be estimable from station measurements. Recording actual population movements during the modeling day, however, seems a nearly unsurmountable task. In reconciling these problems, several options seem available; among these are the following two:

> If the observed concentration field can be estimated acceptably well, both it and the predicted field can be used with the predicted population distribution to compute exposure dosage measures for comparisons. Such a predicted distribution is frequently available when multiple-source, region-oriented issues are being considered. To characterize diurnal variations in emissions, particularly mobile automotive ones, one must estimate the diurnal patterns of population movement. Having done so, one can infer the hourly spatial distribution of population. However, for specific-source issues, population distribution is seldom considered. Since only the emissions from the individual source are of interest, those of the same species resulting from nearly population-related activities need not be explicitly considered, except to compute a background concentration over which the specific-source emissions are superimposed. Unless additional information can be gathered (from a traffic planning agency perhaps), population distribution may not be available, even as a prediction.

> If the observed concentration field is not known acceptably well, computation of the observed exposure/dosage measures cannot be accomplished. However, these quantities often can be

calculated for model predictions (presuming a predicted population distribution history is available). Even though these cannot be compared against their observed values, they can help characterize model predictions. A model sensitivity analysis can be conducted to estimate the effect of population distribution on exposure/dosage calculations. If sensitive, the gathering of additional observational data might be warranted, as would an expanded effort in predicting population movement.

The exposure/dosage performance measures considered here fall into three types: scalar, statistical, and "pattern recognition." We present in Table C-6 some specific measures.

a. Scalar Exposure/Dosage Performance Measures

Several performance measures are defined in terms of concentration exposure and dosage. The exposure is defined to be the product of the number of persons experiencing a concentration in excess of a certain value and the time duration over which the value is exceeded. It is expressed analytically as follows:

$$E^m(x,y,\eta) = \int_{t_1}^{t_2} P(x,y,t)\ u\left[C^m(x,y,t)-\eta\right]dt \quad , \qquad \text{(C-28)}$$

where $E^m(x,y,\eta)$ is the exposure at a point (x,y) to a concentration $C^m(x,y,t)$ of species m in excess of a given level, η (the NAAQS, for example); $P(x,y,t)$ is the population level at (x,y) at time t; u is the unit step function such that

$$u(z) = \begin{cases} 1 \quad , & z \geq 0 \\ 0 \quad , & z < 0 \end{cases} \quad ; \qquad \text{(C-29)}$$

TABLE C-6. SOME EXPOSURE/DOSAGE PERFORMANCE MEASURES

Type	Performance Measure
Scalar	a. Difference for the modeling day in the number of person-hours of exposure to concentrations: 1) Greater than the NAAQS 2) Within 10 percent of the peak. b. Difference for the modeling day in the total pollutant dosage.
Statistical	a. Differences in the exposure/concentration frequency distribution function; differences in the following are of interest: 1) Cumulative distribution function 2) Density function 3) Expected value of concentration 4) Standard deviation of density function b. Cumulative dosage distribution function as a function of time during the modeled day.
Pattern recognition	For each hour during the modeled day, an isopleth plot of the following (both for predictions and observations): 1) Dosage 2) Exposure

and $\Delta t = t_2 - t_1$, is the duration of exposure. The total exposure between t_1 and t_2 over a region measuring X by Y can be written as

$$E^m_T(\eta) = \int_0^Y \int_0^X E^m(x,y,\eta)\,dx\,dy \quad . \tag{C-30}$$

Since in practice the predicted and observed concentration fields are known only at discrete points on a ground-level grid, it follows that the population function $P(x,y,t)$ must be resolved into a compatible, discrete form. Once this is done, the discrete forms of Eqs. (C-28) and (C-30) can be written as follows:

$$E^m_{ij}(\eta) = \sum_{n=N_1}^{N_2} P^n_{ij}\, u\left[C^{m,n}_{ij} - \eta\right] \tag{C-31}$$

$$E^m_T(\eta) = \sum_{j=1}^{J} \sum_{i=1}^{I} E^m_{ij}(\eta) \tag{C-32}$$

where I and J are the X and Y dimensions of the grid while N_1 and N_2 are the starting and ending hours of the summation.

Dosage is defined as the product of the population at a given point, the pollutant concentration to which that population is exposed, and the length of time for which the exposure to that concentration persists. The dosage provides a measure of the total amount of pollutant present in the total volume of air inhaled by people over the time period of interest. This may be illustrated as follows. Let the dosage, D, be in units of ppm-person-hour. If the volume of air inhaled is V cubic meters per person-hour, the quantity of pollutant, Q, present in the air may be estimated as

$$Q = DV \times 10^{-6} \text{ cubic meters} \tag{C-33}$$

If V is assumed to be a constant, then Q is proportional to D and the dosage D provides a measure of Q. It may be noted that the dosage provides no information as to the amount of pollutant inhaled per person. The dosage at a point (x,y) may be expressed as

$$D^m(x,y) = \int_{t_1}^{t_2} P(x,y,t)\, C(x,y,t)dt \qquad \text{(C-34)}$$

while the total dosage within an area X by Y is

$$D_T^m = \int_0^Y \int_0^X D^m(x,y)\, dx\, dy \qquad \text{(C-35)}$$

Expressed in discrete terms these two equations can be written as

$$D_{ij}^m = \sum_{n=N_1}^{N_2} P_{ij}^n\, C_{ij}^{m,n} \qquad \text{(C-36)}$$

$$D_T^m = \sum_{j=1}^{J} \sum_{i=1}^{I} D_{ij}^m \qquad \text{(C-37)}$$

Using Eqs. (C-31) and (C-32) we can calculate two measures of interest: We can determine for the predicted and observed concentrations the number of person-hours of exposure to concentrations (1) greater than the NAAQS and (2) near the peak (within 10 percent, for example). Using Eqs. (C-36) and (C-37), we can determine for the modeling day the total predicted and observed pollutant dosage. By comparison of the predicted and observed values, the seriousness of any differences between the two can be estimated in a way that relates, though crudely, to pollutant health impact.

b. Statistical Exposure/Dosage Performance Measures

Exposure/dosage performance measures have several useful statistical variants. One of these is the difference between the predicted and observed exposure/concentration distribution function. An example of such a function is shown in Figure C-16, calculated for ozone in the Denver Metropolitan

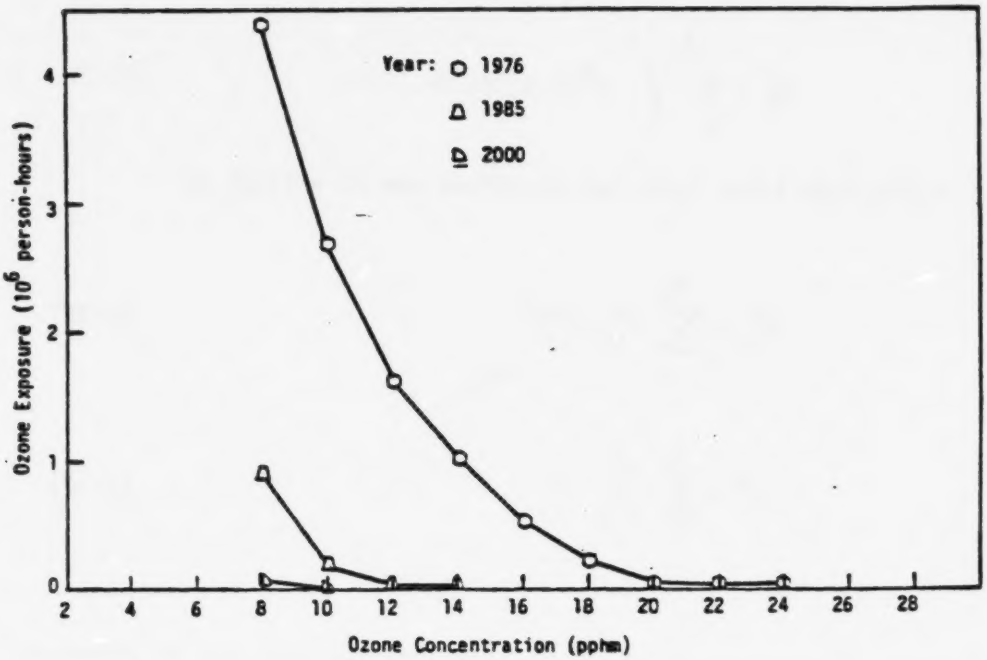


FIGURE C-16. ESTIMATED EXPOSURE TO OZONE AS A FUNCTION OF OZONE CONCENTRATION FOR 3 AUGUST 1976 METEOROLOGY. This figure is based on predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.

region. The figure is based on predictions made by the SAI Urban Airshed Model using actual emissions and meteorology for 3 August 1976, as well as projected emissions for 1985 and 2000.

Certain statistics of the exposure distribution are useful: the cumulative distribution function (CDF) itself, the density function (f_E), the expected value of the pollutant concentration, and the standard deviation of the density function. We show in Figure C-17 a representation of the general shapes taken by the CDF_E and the f_E.

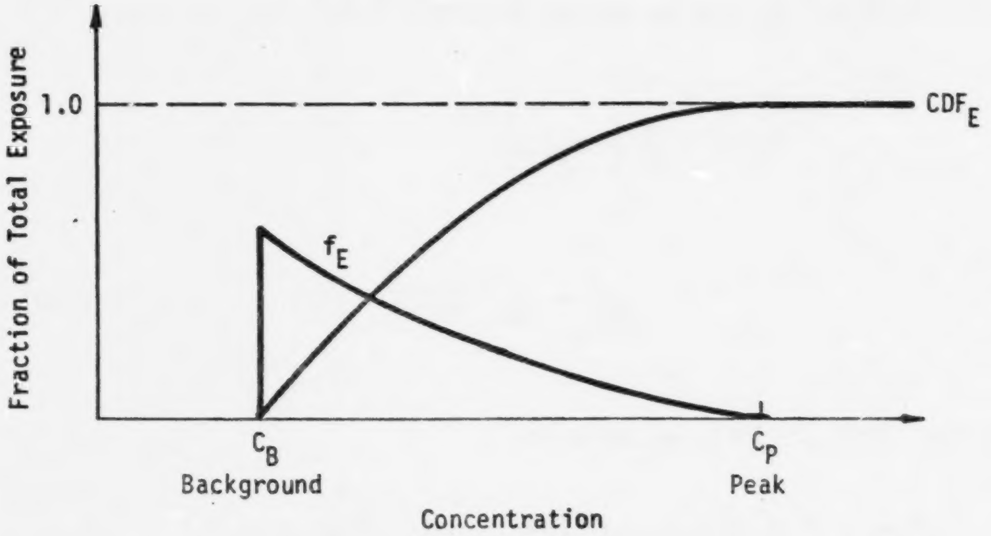


FIGURE C-17. GENERAL SHAPE OF THE EXPOSURE CUMULATIVE DISTRIBUTION AND DENSITY FUNCTIONS

Incorporated in this figure are two important assumptions: None of the population is exposed to concentrations above the peak value, C_P, while all are exposed to concentrations at least as high as the background value, C_B. The first of these is certainly a valid assumption. The second may not be accurate in all circumstances. Those persons spending their days indoors within environmentally controlled buildings may experience lesser concentrations than the background value. Noting this possible limitation, however, we proceed.

The CDF_E can be derived from the exposure function defined in Eq. (C-30) and illustrated with the example in Figure C-16. It can be expressed as

$$CDF_E(C) = 1 - \frac{E_T^m(C)}{E_T^m(C_B)} \tag{C-38}$$

The density function, f_E, can be derived from this relation as follows:

$$f_E(C) = \frac{d}{dC}\left[CDF_E(C)\right]$$

$$= -\frac{1}{E_T^m(C_B)}\,\frac{d}{dC}\left[E_T^m(C)\right] \tag{C-39}$$

Combining Eqs. (29) and (31), we can write

$$E_T^m(C) = \int_Y\int_X\int_t P(x,y,t)\; u\left[C^m(x,y,t) - C\right] dt\, dx\, dy \tag{C-40}$$

From this, we can express its derivative as

$$\frac{d}{dC}\left[E_T^m(C)\right] = -\int_Y\int_X\int_t P(x,y,t)\;\delta\left[C^m(x,y,t) - C\right] dt\, dx\, dy \tag{C-41}$$

where δ is the Dirac delta function defined such that $\delta(z)$ is 1 when $z = 0$ and zero for all other values of z. The density function can thus be written as

$$f_E(C) = \frac{1}{E_T^m(C_B)} \int_Y \int_X \int_t P(x,y,t)\delta\left[C^m(x,y,t) - C\right] dt\ dx\ dy \qquad \text{(C-42)}$$

The expected value, μ_E, and the standard deviation, σ_E, are defined as follows:

$$\mu_E = \int_{C_B}^{C_P} Cf_E(C)\ dC \qquad \text{(C-43)}$$

$$\sigma_E^2 = \int_{C_B}^{C_P} (C - \mu_E)^2 f_E(C)\ dC \qquad \text{(C-44)}$$

Because the concentration field and population distribution usually are not known continuously but only at a set of fixed points, the discrete forms of the above equations are usually of greater practical value. The CDF_E remains as expressed in Eq. (C-38) but the exposure quantities must be calculated using Eqs. (C-31) and (C-32). The density function, however, involves the use of the delta function. With a discretely expressed concentration field, the argument of $\delta(z)$ may seldom be zero. To overcome this problem, we approximate the delta function with an expression that remains constant over a small interval, ΔC, about $C_{ij}^{m,n}$, the discrete form of the concentration $C^m(x,y,t)$. This approximation can be expressed as

$$\psi\left[C_{ij}^{m,n} - C\right] = u\left(C_{ij}^{m,n} - C + \frac{\Delta C}{2}\right) - u\left(C_{ij}^{m,n} - C - \frac{\Delta C}{2}\right) \qquad \text{(C-45)}$$

This function has the form shown in Figure C-18.

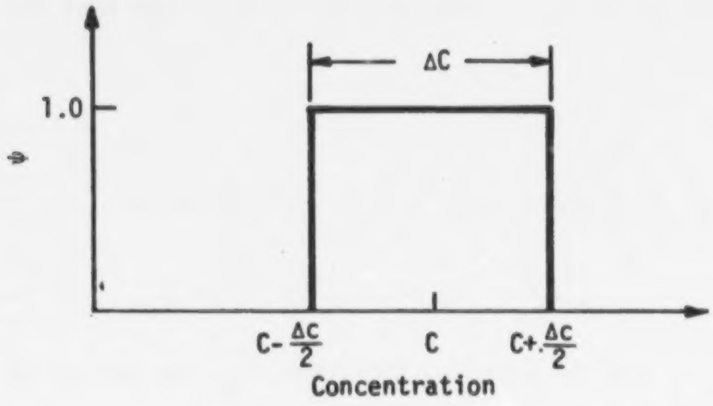


FIGURE C-18. SHAPE OF ψ(C), THE APPROXIMATION TO THE DELTA FUNCTION

Using Eq. (C-45) the discrete form of the density function can be written in the following form:

$$f_E(C) \cong \frac{1}{E_T^m(C_B)} \sum_{j=1}^{J} \sum_{i=1}^{I} \sum_{n=N_1}^{N_2} P_{ij}^n \, \psi\left[C_{ij}^{m,n} - C\right] \tag{C-46}$$

The expected value and standard deviation then can be expressed as

$$\mu_E = \frac{1}{K} \sum_{k=1}^{K} C_k f_E(C_k) \tag{C-47}$$

$$\sigma_E^2 = \frac{1}{K - 1} \sum_{k=1}^{K} (C_k - \mu_E)^2 \, f_E(C_k) \tag{C-48}$$

where K is the number of equally spaced intervals, ΔC, spanning the concentration range from C_B to C_P.

The quantities described above--the CDF_E, f_E, μ_E and σ_E--form the basis for a comparison between prediction and observation. Differences in the shape of the CDF_E can be characterized by differences in μ_E and σ_E^2, as well as being revealed by differences in the qualitative shapes of the f_E. If these differences are large, model performance may be judged unacceptable.

The variation of the cumulative dosage function during the modeling day is another means for comparing prediction with observation. An example of such a dosage function is shown in Figure C-19, calculated for ozone in the Denver Metropolitan region. The figure is based on predictions made by the SAI Urban Airshed Model.

c. "Pattern Recognition" Exposure/Dosage Performance Measures

The performance of a model in predicting exposure and dosage can be judged qualitatively by comparing isopleth plots of predicted values with a similar plot showing observed ones. We present in Figures C-20 and C-21 the ozone exposure and dosage contours, respectively, predicted by the SAI Urban Airshed Model for Denver on 3 August 1976. The population distribution assumed in each was based on data supplied by the Denver Regional Council of Governments. Residential population figures were corrected temporally to account for daytime employment patterns. No attempt was made, however, to adjust for other shifts during the day.

In Figure C-20, the cumulative exposure at one-mile intervals is shown. Isopleths of exposure to concentrations greater than a certain value are included for three different levels. In Figure C-21, the cumulative dosages are shown for each point on the same one-mile spaced grid. In both figures, the interval of time considered was 13 hours, from 500 to 1800 (MST).

5. "HYBRID" PERFORMANCE MEASURES

As noted earlier, model predictions often are more finely resolved spatially than measurement data. A consequence of this is the following:

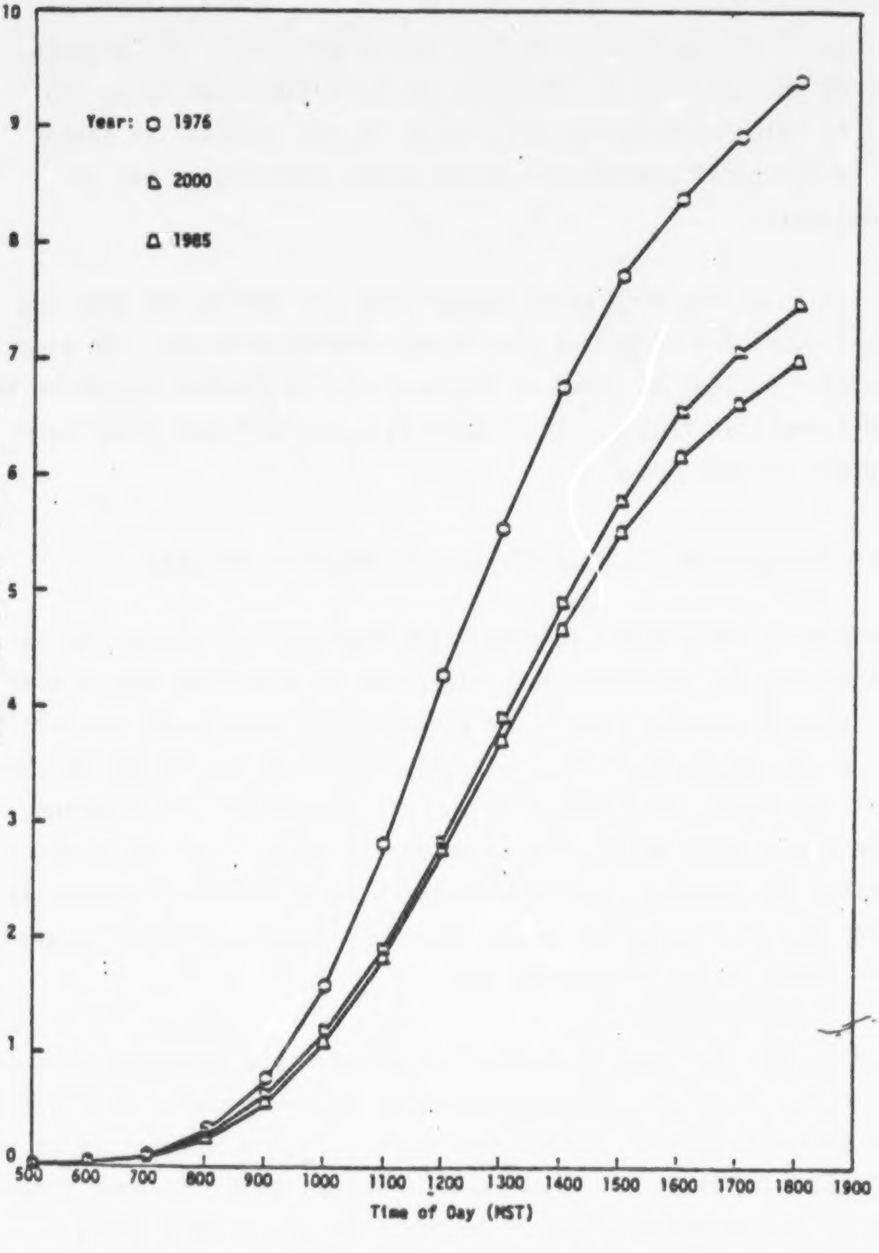
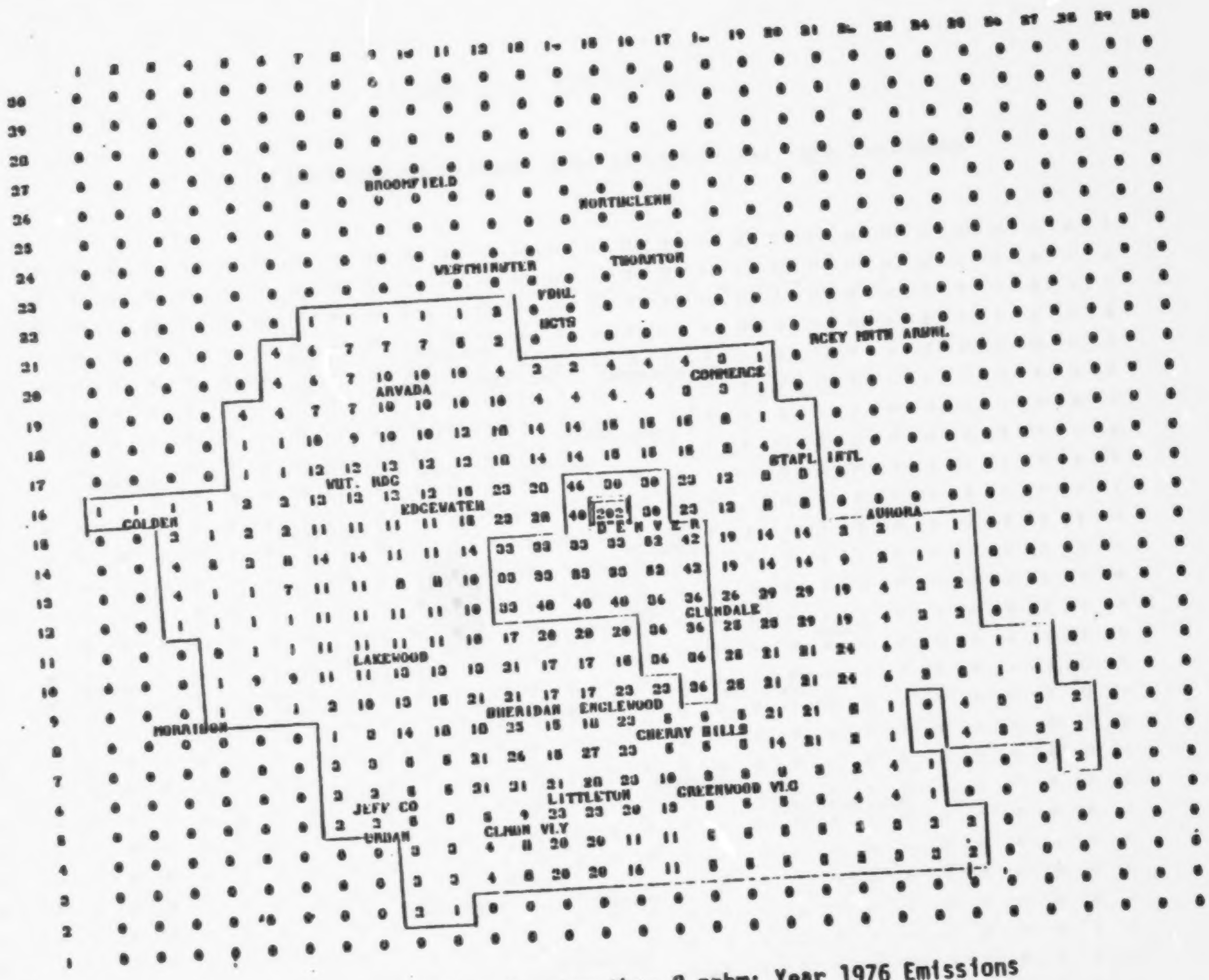
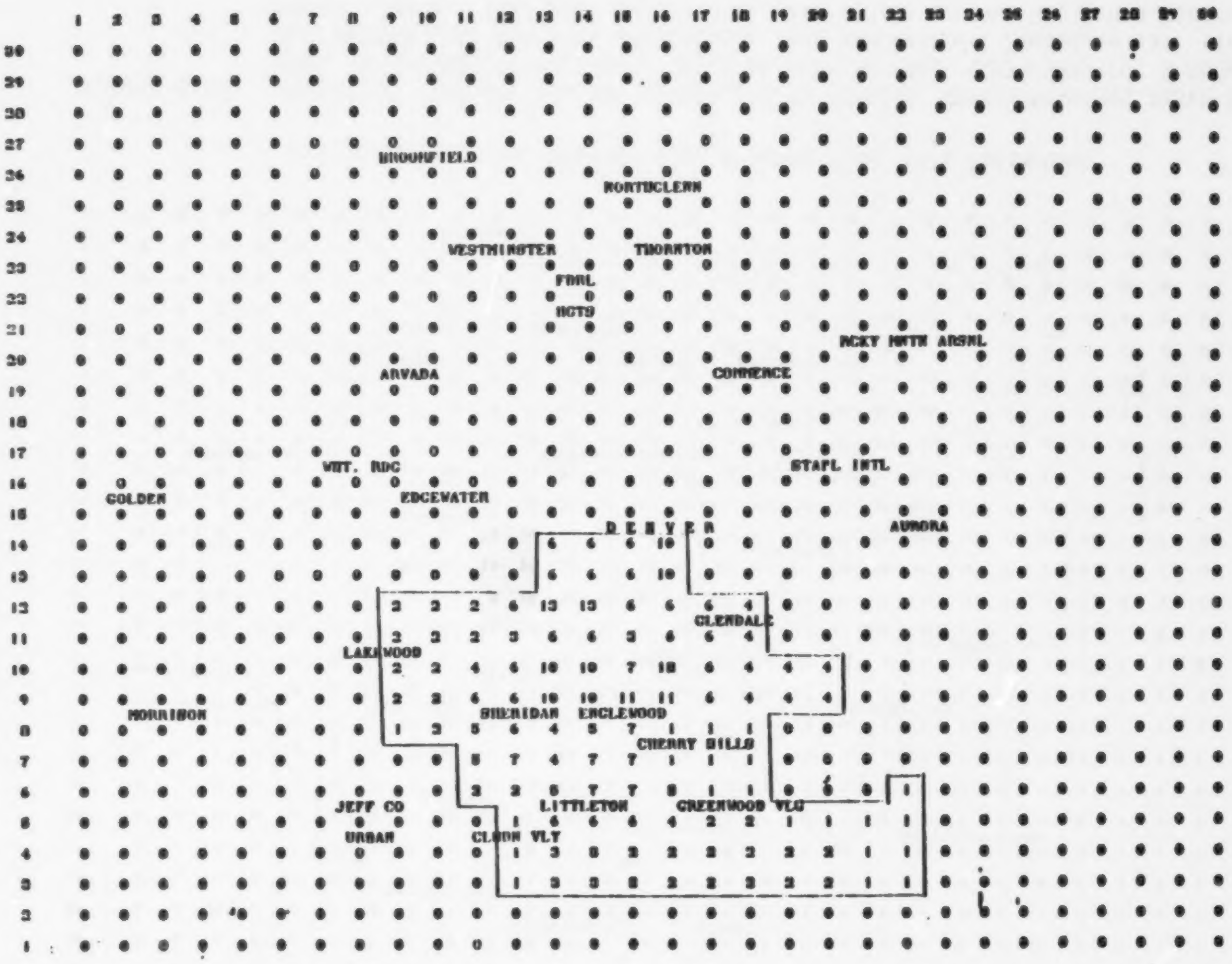


FIGURE C-19. CUMULATIVE OZONE DOSAGE AS A FUNCTION OF TIME OF DAY FOR 3 AUGUST 1976 METEOROLOGY. This figure is based on the predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.


BROOMFIELD
NORTHGLENN
WESTMINSTER
THORNTON
FDRL HGTS
RCKY MNTN ARSNL
COMMERCE
ARVADA
STAPL INTL
WHT. RDG
EDGEWATER
AURORA
GOLDEN
DENVER
GLENDALE
LAKEWOOD
SHERIDAN
ENGLEWOOD
MORRISON
CHERRY HILLS
LITTLETON
GREENWOOD VLG
JEFF CO
URBAN
CLMBN VLY

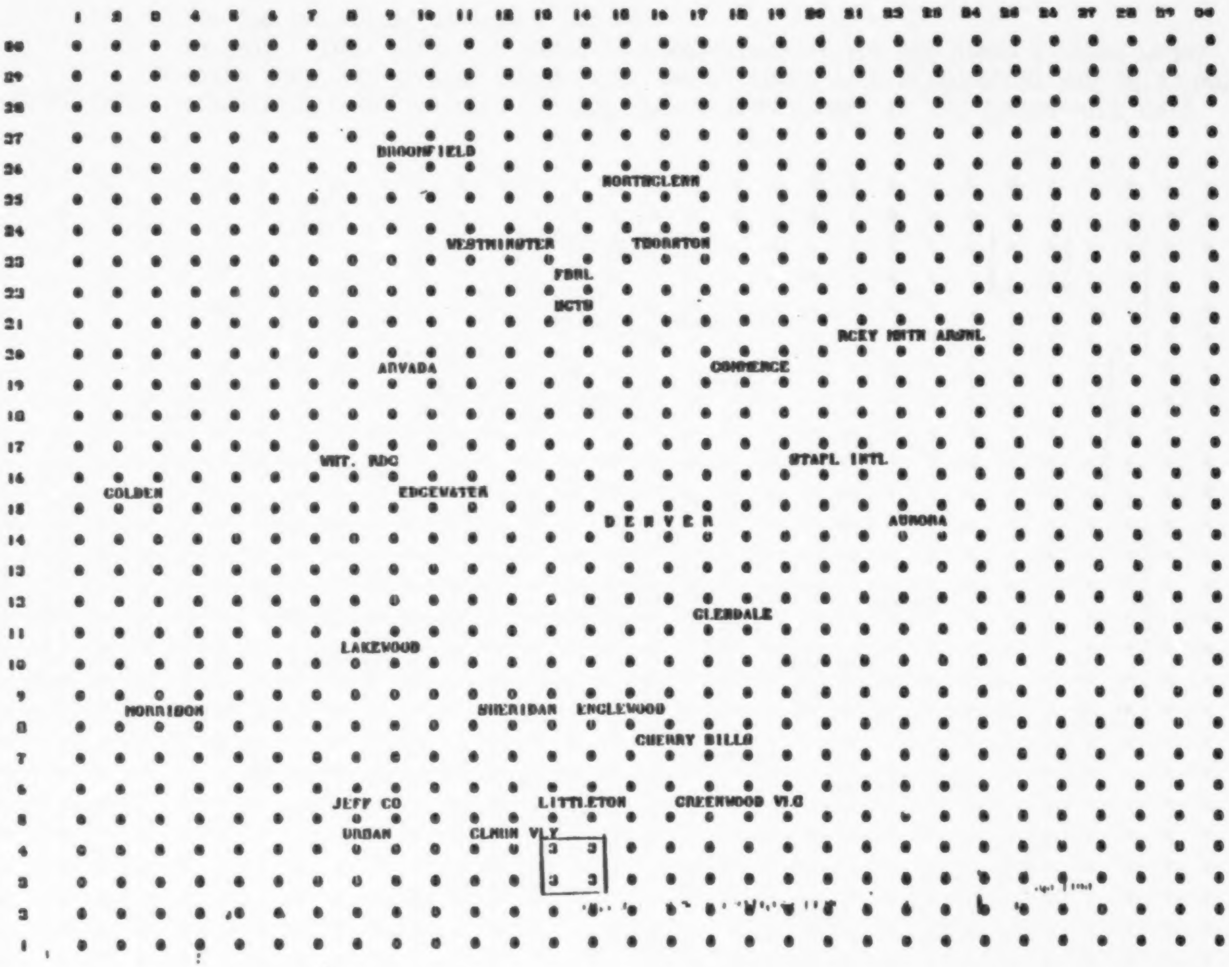
(a) Concentration Greater than 8 pphm; Year 1976 Emissions

FIGURE C-20. CUMULATIVE EXPOSURE (IN 10^3 PERSON-HOURS) TO OZONE CONCENTRATIONS ABOVE GIVEN LEVEL IN ONE-SQUARE-MILE GRID CELLS BETWEEN 500 AND 1800 HOURS FOR 3 AUGUST 1976 METEOROLOGY AND 1976 EMISSIONS. Grid numbers are listed on left side and top of figure. This plot is based on predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.


BROOMFIELD
NORTHGLENN
WESTMINSTER
THORNTON
FDRL
HGTS
RCKY MNTN ARSNL
ARVADA
COMMERCE
WHT. RDG
STAPL INTL
GOLDEN
EDGEWATER
DENVER
AURORA
GLENDALE
LAKEWOOD
MORRISON
SHERIDAN
ENGLEWOOD
CHERRY HILLS
LITTLETON
GREENWOOD VLG
JEFF CO
URBAN
CLRDN VLY

(b) Concentration Greater than 16 pphm; Year 1976 Emissions

FIGURE C-20 (Continued)


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
BROOMFIELD
NORTHGLENN
WESTMINSTER
THORNTON
FBRL
DCTD
RCKY MNTN ARSNL
ARVADA
COMMERCE
WHT. RDG
STAPL INTL
GOLDEN
EDGEWATER
D E N V E R
AURORA
GLENDALE
LAKEWOOD
MORRISON
SHERIDAN
ENGLEWOOD
CHERRY HILLS
JEFF CO
LITTLETON
GREENWOOD VLG
URBAN
CLMBN VLY
3
3
3
3

(c) Concentration Greater than 24 pphm; Year 1976 Emissions

FIGURE C-20 (Concluded)

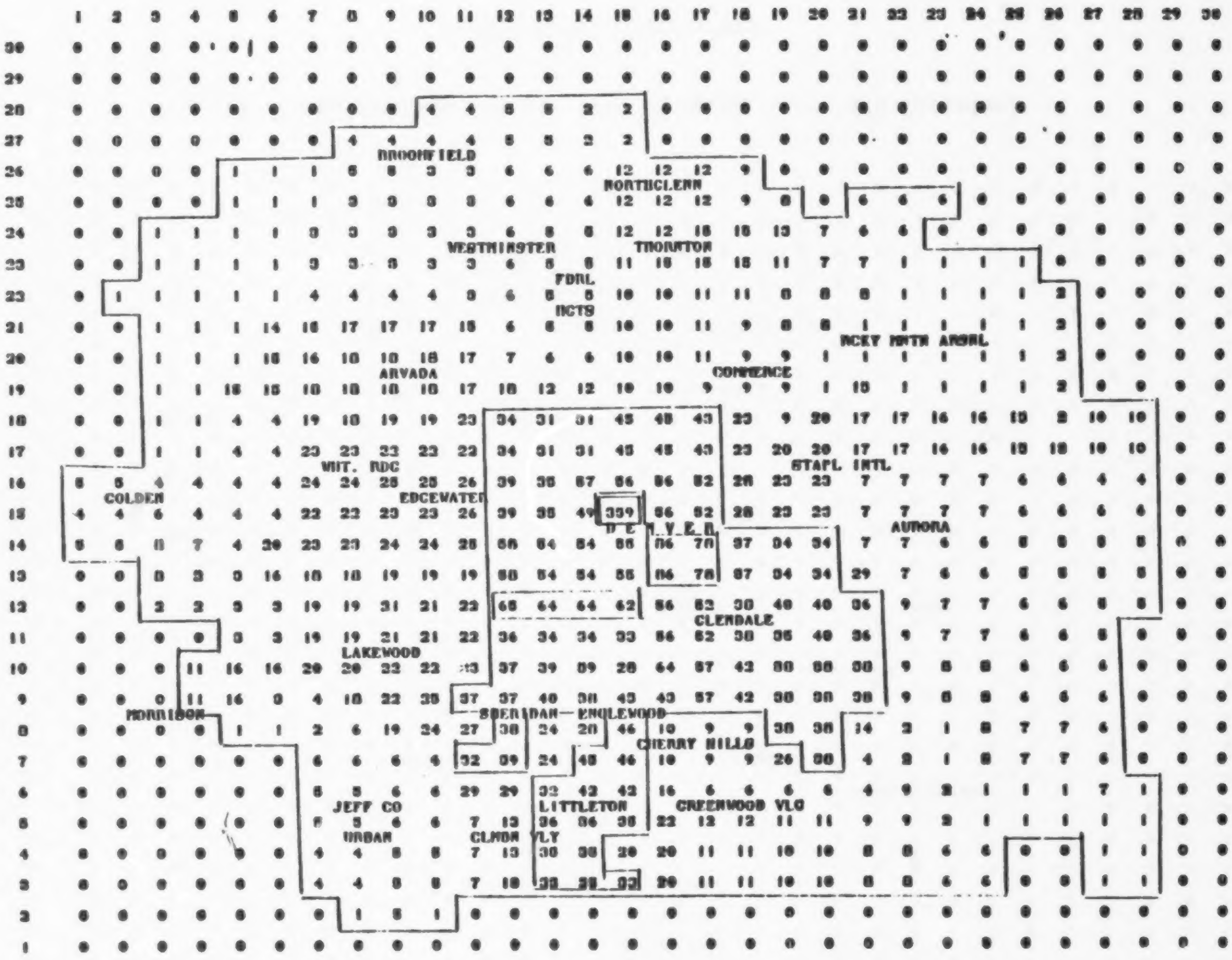

BROOMFIELD
NORTHGLENN
WESTMINSTER
THORNTON
FDRL
HCTS
RCKY MNTN ARSNL
ARVADA
COMMERCE
WHT. RDG
STAPL INTL
GOLDEN
EDGEWATER
DENVER
AURORA
GLENDALE
LAKEWOOD
MORRISON
SHERIDAN—ENGLEWOOD
CHERRY HILLS
LITTLETON
GREENWOOD VLG
JEFF CO
URBAN
CLMBN VLY

FIGURE C-21. CUMULATIVE OZONE DOSAGES (IN 10^6 PPHM-PERSON-HOURS) IN ONE-SQUARE-MILE GRID CELLS FROM 500 TO 1800 HOURS (MST) for 3 AUGUST 1976 METEOROLOGY AND EMISSIONS IN 1976. This figure is based on predictions of the SAI Urban Airshed Model for the Denver Metropolitan region.

model performance sometimes must be evaluated using performance measures requiring different classes of data "completeness." For instance, the observed concentration field may not be inferred reliably from station data even though the predicted field can be well described. In such a case, concentration isopleth plots for both could not be constructed and compared directly. Still, we would not wish to rely solely on station performance measures. To do so, we would sacrifice some of the information content available on the prediction side of the comparison.

Several performance measures are "hybrid" ones. They are designed for use when a different level of concentration information is available for prediction than for observation. We discuss here such a measure, the basis for which is shown in Figure C-22.

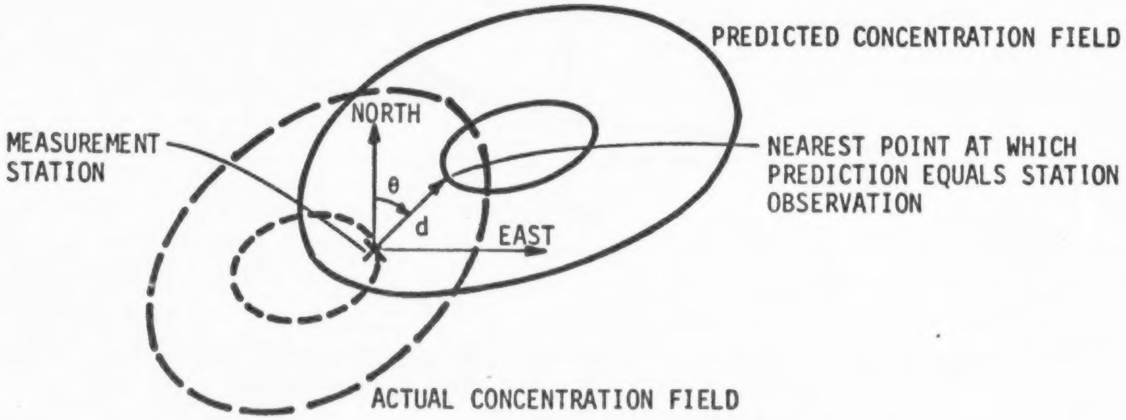

PREDICTED CONCENTRATION FIELD
NORTH
MEASUREMENT STATION
θ
d
EAST
NEAREST POINT AT WHICH PREDICTION EQUALS STATION OBSERVATION
ACTUAL CONCENTRATION FIELD

FIGURE C-22. ORIENTATION WITH RESPECT TO MEASUREMENT STATION OF NEAREST POINT AT WHICH PREDICTION EQUALS STATION OBSERVATION

In the figure, isopleths are shown for the predicted and actual concentration fields. Only at the measurement station, however, is data available describing the actual field. The offset between the two fields nevertheless can be characterized by determining the vector (distance, azimuthal orientation) from the station to the nearest point at which the predicted concentration equals the measured value. This can be done for several hours, producing a time history of the distance and orientation of that point. A plot of this can be constructed, as shown in Figure C-23.

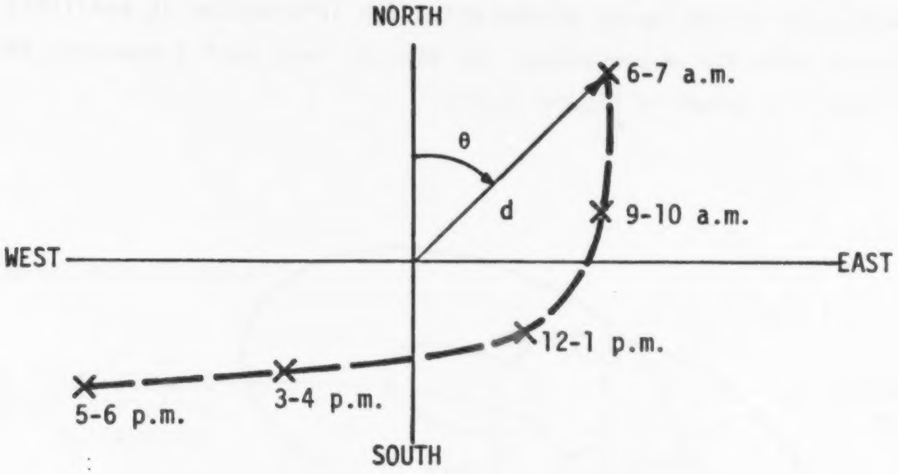


FIGURE C-23. SPACE-TIME TRACE OF LOCATION OF NEAREST POINT PREDICTING A CONCENTRATION EQUAL TO THE STATION MEASURED VALUE

The space-time trace shown in the figure is centered at the measurement station. Similar traces could be constructed for each station. Space-time correlations could be made to infer the amount and orientation of the displacement of the two concentration fields.

APPENDIX D

SEVERAL RATIONALES FOR SETTING MODEL PERFORMANCE STANDARDS

APPENDIX D

SEVERAL RATIONALES FOR SETTING MODEL PERFORMANCE STANDARDS

In Chapter VI of this report, we identify a "preferred" set of model performance measures, the values of which are helpful in assessing the degree to which model predictions agree with observations. It remains for us to decide how "close" these must be in order to judge model performance to be acceptably good. In this appendix, we present four alternate rationales for making such decisions: Health Effects, Control Level Uncertainty, Guaranteed Compliance, and Pragmatic Historic. To maintain perspective about each rationale and the problems for which their use may be appropriate, we recommend Section D of Chapter VI be read prior to considering this appendix.

1. Health Effects Rationale

Ambient pollutant concentrations are not themselves our most fundamental concern but rather the adverse health effects they produce. The NAAQS are chosen to serve as measurable, enforceable surrogates for the "acceptable" levels of health impact they imply. Because health effects are of such basic importance, it makes sense to define model performance in such terms. However, quantifying the health effects resulting from exposure to a specified pollutant level can be a difficult and controversial task. Toxicological studies in laboratories by necessity are performed at high concentrations, often at levels and dosages seldom occuring even in the most polluted urban areas. Experiments are conducted in animals whose response patterns may not serve as perfect analogues for human behavior. Epidemiological studies are confounded by the variety of effects occuring simultaneously in a complex urban environment. Consequently, isolation of a "cause-and-effect" relationship between health effect and pollutant level becomes statistically very difficult.

Nevertheless, in this discussion we indicate one means whereby health effects can be used as a basis for evaluating the acceptability of model performance. We postulate the existence of a health effects functional, ϕ, dependent both on concentration level and health effects for all exposed persons in the polluted region. This quantity (the area-integrated cumulative health effect) we use as the metric of interest. If the ratio of the predicted value of ϕ to its observed value remains within a certain tolerance of unity, model performance is judged acceptable.

Several features of this approach have appeal. Among these are:

- > The health effects functional need not be known precisely, only its general shape.
- > The use of area-integrated cumulative health effects as a metric has strong intuitive appeal; it is less sensitive than dosage to concentrations not near the peak value.
- > A transformation of variables reduces the spatial sensitivity of the metric, ϕ, with more than one spatially distributed region mapped in to the same value of ϕ; this can result in an increase in generality of application.
- > Simplifying assumptions can be invoked to allow computation of specific numerical values.

a. Area Cumulative Health Effects As a Concept

"Total area dosage" is frequently used as a surrogate for "total area health effects." Mathematically, total area dosage, D_T can be expressed as

$$D_T(t_1,t_2) = \int_X \int_Y \int_{t_1}^{t_2} P(x,y,t)C(x,y,t)dt\ dy\ dx \qquad \text{(D-1)}$$

where the duration of exposure is Δt $(=t_2-t_1)$; $P(x,y,t)$ and $C(x,y,t)$ are the population and concentration at (x,y) at time t; and X and Y represent the spatial limits of the polluted region.

However, the concentration $C(x,y,t)$ in this relation and the time duration of exposure really combine to approximate health effects. Suppose that a health effects function exists such that

$$HE = HE(C,\Delta t) \quad . \qquad \text{(D-2)}$$

Such a function could behave as shown in Figure D-1, with HE disappearing only when concentrations approach zero. Alternatively, a threshold concentration might exist below which specific effects are either indistinguishable from a background level or below the threshold of perception.

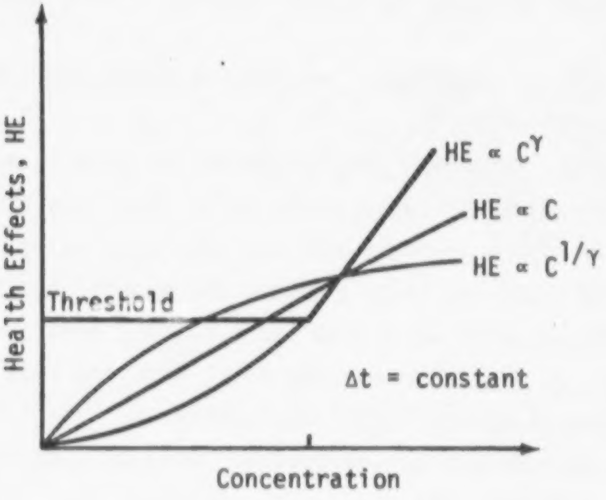


FIGURE D-1. POSSIBLE HEALTH EFFECTS CURVES

We define a new metric: the area-integrated cumulative health effects functional, ϕ. It can be written as follows:

$$\phi(\Delta t) = \int_X \int_Y \int_{\Delta t} P(x,y,t)\ HE[C(x,y,t),t-t_1]dt\ dy\ dx \qquad \text{(D-3)}$$

If this function could be evaluated for predicted and "true" values of P(x,y,t) and C(x,y,t), we could formulate the performance standard such that their ratio, r, was required to remain within a fixed tolerance of unity, i.e.,

$$r = \frac{\phi(\Delta t)|_{\text{predicted}}}{\phi(\Delta t)|_{\text{observed}}} \geq 1 - \alpha \qquad \text{(D-4)}$$

where α is some small value (10 percent, for instance) chosen to represent a maximum acceptable level of uncertainty in aggregate health impact. It may be noted with this standard that model acceptability is called into doubt only if the predicted value of ϕ is less than the "observed" value. This makes sense for the following reason: Considering only a perspective based on health effects, we are concerned that the model predict conditions leading to health impact at least (or nearly so) as large as actually occurs. To bound model on the "upper" side, another rationale must be used (control level uncertainty, perhaps).

The expression in Eq. D-3, however, is of only academic interest unless it can be made more tractable. Several of its key limitations are as follows:

> It is a spatial integral. The value of $P(x,y,t)$ and $C(x,y,t)$ change for each new application locale. Thus it is difficult to extend results obtained in one situation to those expected in any new one.
> The health effects function, HE, is dependent on concentration and cannot be expressed directly without being "mapped" through the concentration field.

However, through a transformation of variables, some difficulties can be overcome. We will replace in Eq. D-3 the double spatial integration by a single concentration integration taken over the range of ambient values (background, C_B, to the current peak, C_P). Total population within the modeling region at time t, $P_T(t)$, can be written as

$$\int_X\int_Y P(x,y,t)\,dydx = P_T(t) = \int_{C_B}^{C_P(t)} w(C,t)\,dC \qquad \text{(D-5)}$$

where $w(C,t)$ is the population exposed to a concentration C at time t. (By definition, no one is exposed to concentrations lower than the background value, C_B.) A pictorial representation of the population function $P(x,y,t)$ and $w(C,t)$ is shown in Figure D-2.

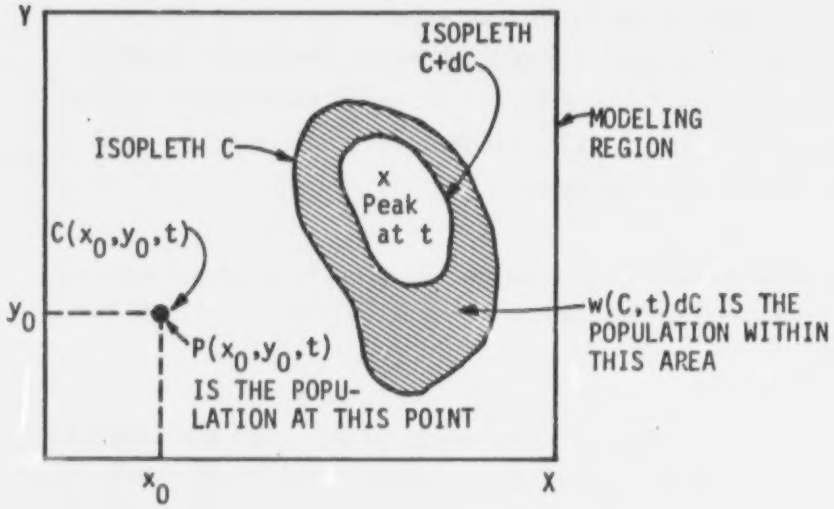

Y
ISOPLETH C+dC
MODELING REGION
ISOPLETH C
x Peak at t
$C(x_0,y_0,t)$
y_0
$P(x_0,y_0,t)$ IS THE POPULATION AT THIS POINT
w(C,t)dC IS THE POPULATION WITHIN THIS AREA
x_0
X

FIGURE D-2. REPRESENTATION OF SPATIAL AND CONCENTRATION DEPENDENT POPULATION FUNCTIONS

The equivalence expressed in Eq. D-5 holds without qualification providing the modeling region is chosen large enough to contain the background (C_B) isopleth for every hour during the day. However, this requirement can be relaxed under the following condition: No or very few persons live or work in the area outside the modeling region but within the C_B isopleth. In such a case the modeling region need only be large enough to enclose within it the population of interest.

An important observation can now be made: The health effects function, HE, can be introduced into both sides of Eq. D-5 without disturbing the equality. Doing so and integrating with respect to time, the area integrated cumulative health effects (CHE) functional can be transformed into

$$\phi(\Delta t) = \int_{\Delta t}\int_{C_B}^{C_P(t)} w(C,t)HE(C,t-t_1)\ dCdt \qquad (D-6)$$

It is this equation with which we deal in the remainder of this section.

b. Components of the Cumulative Health Effects Functional

We now examine each of the two major components of the CHE functional: the population distribution and health effects function. For Eq. D-6 to be of any use to us, it must be made analytic in a way that has a degree of generality from one application locale to another. Consequently, we are guided by three principal objectives: Both $W(C,t)$ and $HE(C,\Delta t)$ must be analytic, integrable, and based upon simple, easily understood assumptions. To accomplish this, important simplifications are invoked. The degree to which they limit the generality of the results is discussed, although additional research beyond the scope of this study seems desirable.

Population Distribution Function

The function $w(C,t)$ represents the distribution of population with respect to both concentration level and time of day. As a first approximation, we assume it is separable, i.e.,

$$w(C,t) = \overline{w}(c)\ f_W(t) \quad , \qquad (D-7)$$

where $\overline{w}(C)$ is the distribution of daytime (workday) population with respect to concentration level alone at a particular fixed time (the time of the concentration peak, for example), and $f_W(t)$ is a weighting function chosen to reflect the diurnal variation in that distribution (residential vs. commute vs. work hours).

Within a pollutant cloud, concentrations tend to be distributed as follows: A distinct peak value occurs, with concentration falling off as a function of radial distance from that peak. Contours of constant concentration (isopleth lines) surround the peak concentrically, with concentration diminishing to background levels. This radial distribution of concentration level is suggestive. If population is distributed about the peak such that

$$P(C) = \int_0^{2\pi} \int_{r(C)}^{0} p(r^*,\theta) r^* dr^* d\theta \quad , \qquad \text{(D-8)}$$

then we can write the following relation:

$$\begin{aligned} \overline{w}(C) &= \frac{dP(C)}{dC} \\ &= \frac{dP}{dr} \frac{dr}{dC} \\ &= -r \frac{dr}{dC} \int_0^{2\pi} p(r,\theta) d\theta \quad , \qquad \text{(D-9)} \end{aligned}$$

where r(C) is the radial distance from the peak at which a concentration C is experienced and $p(r,\theta)$ is the population density at a point located with respect to the peak by its radial and azimuthal polar coordinates, r and θ.

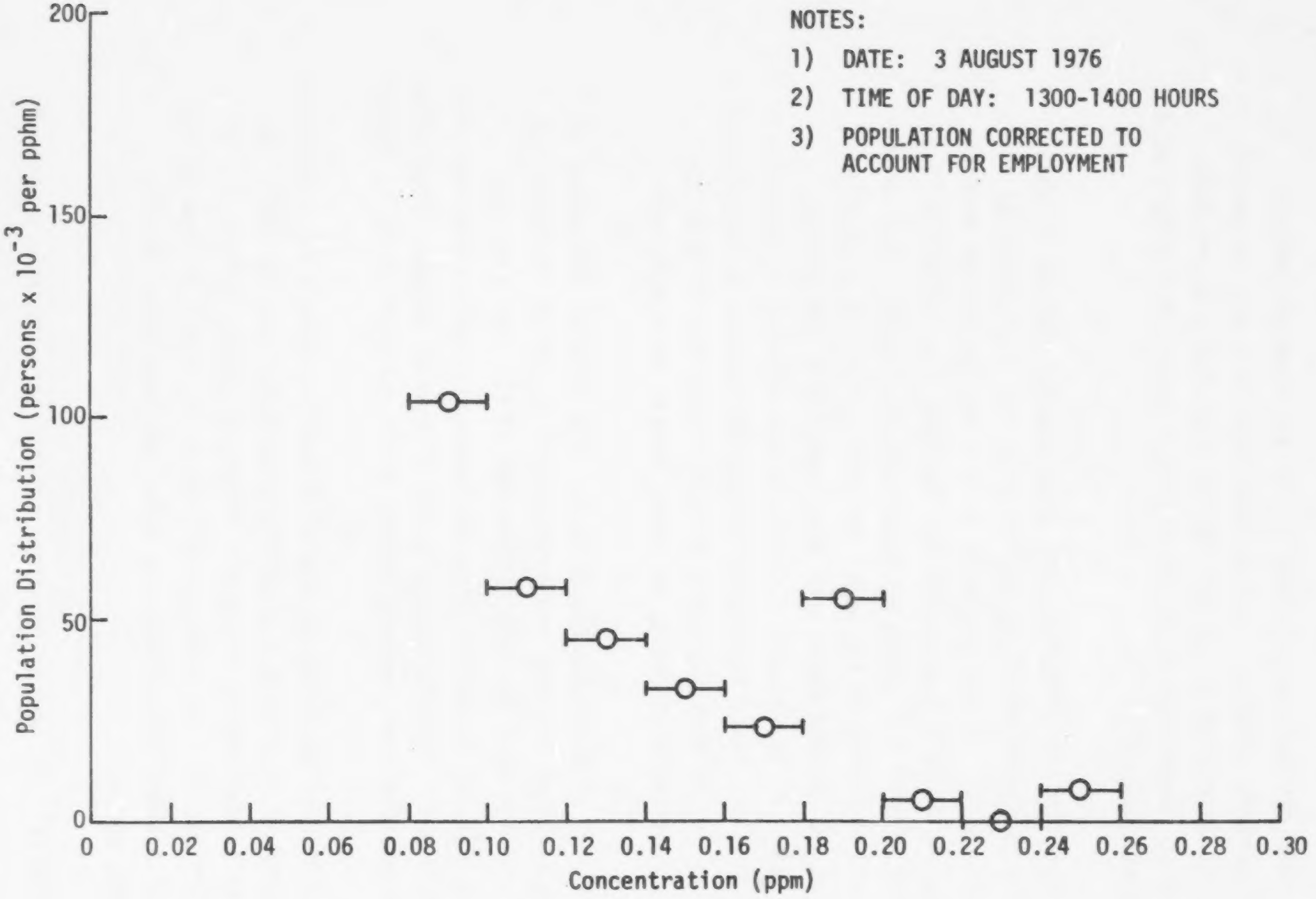
At first glance this formulation would seem neither analytic nor general. The shape of isopleth contours and thus r(C)-- differs considerably from one application to another, even from one hour to the next. The population distribution also would seem highly application-dependent. Further, for reactive species, by the time the peak occurs the pollutant cloud may have drifted some distance (10-30 km) from the densest population centers. However, our approach here is highly pragmatic. To render Eq. D-9 soluble, we must invoke simplifying assumptions. Having done so, comparison of our results with actual data offers us a measure of our success.

Such data has been obtained from ozone exposure/dosage studies done for the Denver Metropolitan region using the grid-based SAI Urban Airshed Model. Shown in Figure D-3 is the population density function predicted on 3 August 1976 for the hour from 1300-1400 (1 to 2 p.m.)--the time of the predicted ozone peak (0.24 ppm). The concentration field predicted by the model was used. A coarse population distribution was derived based upon data supplied by the Denver Regional Council of Governments (DRCOG) and was adjusted to approximate employment shifts. Since the analysis supplied exposure estimates only above 0.08 ppm which were expressed no more finely than in 0.02 ppm increments, an uncertainty band, as shown, exists about each point.

Several key observations can be made. The value of $\overline{w}(C)$ seems to become very small at the peak concentration, i.e., while concentration levels may be high near the peak (within 90% of it), the area (and population) affected is small. Also, an apparent anomaly occurs between 0.18 and 0.20 ppm. This may be due to any of several causes. Population density non-uniformities, however, appear to be the most likely of these.

Using the data contained in Figure D-3 as a standard for comparison, we may proceed in developing a simplified, analytic form for $\overline{w}(C)$. We make two key assumptions in doing so. First, we assume a shape for the radial concentration distribution, $C(r)$, which we invert to give us $r(C)$. Then we make a simplifying assumption about the population density distribution, $p(r,\theta)$.

To estimate $C(r)$, we may idealize isopleth contours as a series of concentric circles, as shown in Figure D-4. Further, we may assume



NOTES:

1) DATE: 3 AUGUST 1976
2) TIME OF DAY: 1300-1400 HOURS
3) POPULATION CORRECTED TO ACCOUNT FOR EMPLOYMENT

FIGURE D-3. POPULATION DISTRIBUTION AS A FUNCTION OF CONCENTRATIONS. Based on predictions of the SAI Urban Airshed Model for the Denver metropolitan region.

there to be N isopleths between the peak concentration, C_P, and the background value, C_B.

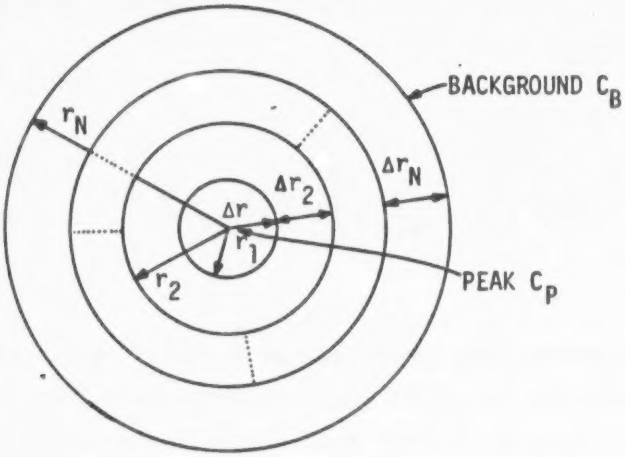


FIGURE D-4. IDEALIZED CONCENTRATION ISOPLETHS

If we assume that for isopleths separated by a constant concentration decrement, ΔC, the interisopleth distance grows exponentially (that is, the isopleths are separated by a steadily growing distance), then we may write an expression for the n-th radius such that

$$r_n = \sum_{i=1}^{n} \Delta r_i$$

$$= \Delta r \sum_{i=0}^{n-1} e^{bi}$$

$$= \left(\frac{1 - e^{bn}}{1 - e^{b}}\right) \Delta r \qquad \text{(D-10)}$$

Since

$$C_n = C_P - n\left(\frac{C_P - C_B}{N}\right) \quad , \tag{D-11}$$

we can solve for n, substitute this into Eq. D-10, and then generalize to yield the following:

$$C(r) = C_P - \frac{\Delta C}{b} \ln\left[1 - \left(\frac{1 - e^b}{\Delta r}\right)r\right] \quad , \tag{D-12}$$

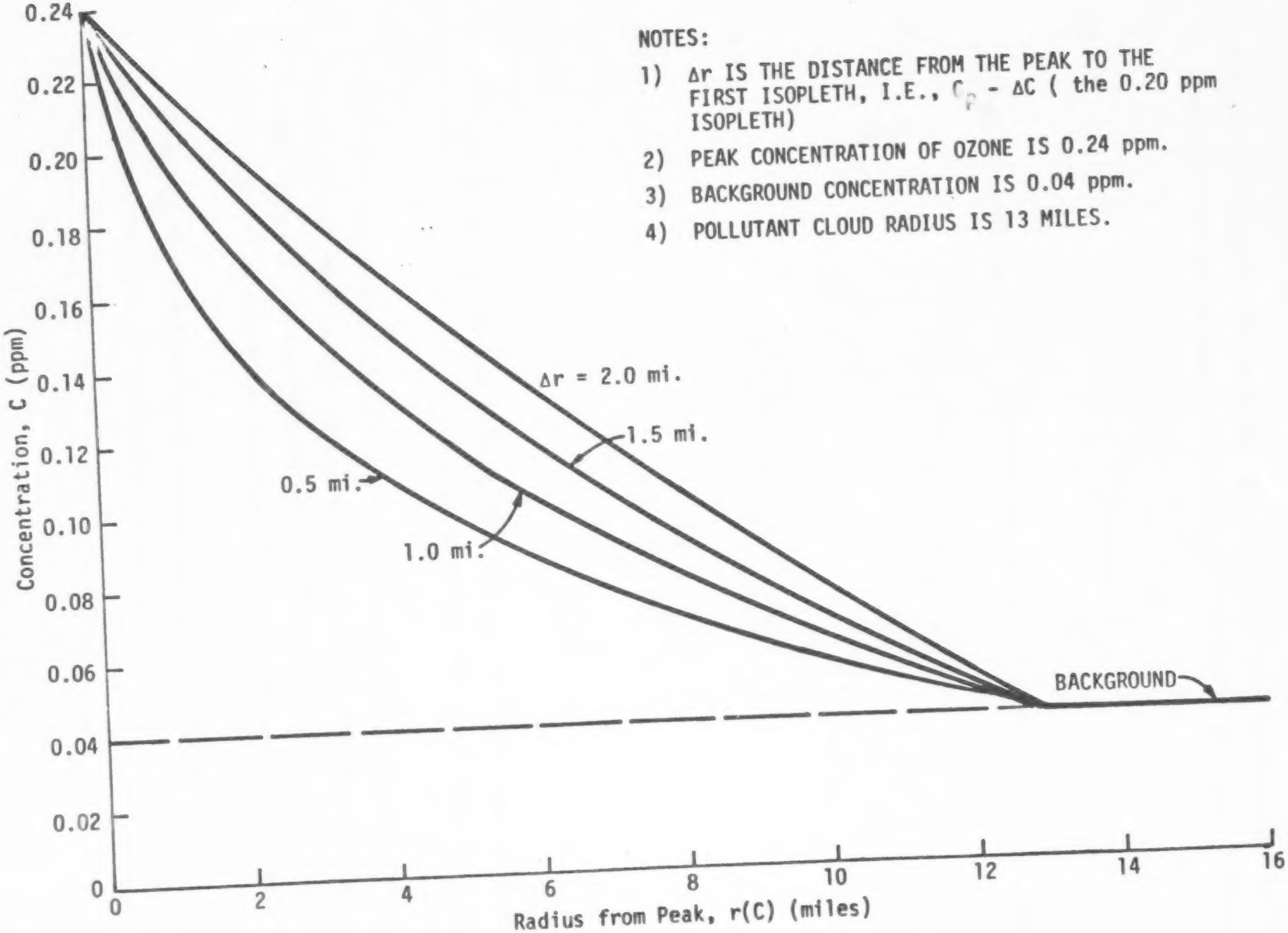
where ΔC is the interisopleth concentration decrement and b is chosen so that $r(C_B)$ equals the radius of the pollutant cloud (here assumed to be the urban radius). Several typical such concentration distributions are shown in Figure D-5.

We can now invert this relation to estimate r(C). Doing so, we can write

$$\begin{aligned} r(C) &= \left(\frac{\Delta r}{1 - e^b}\right)\left[1 - \exp\left(\frac{C_P - C}{\Delta C/b}\right)\right] \\ &= K_1\left(1 - K_2 e^{-K_3 C}\right) \end{aligned} \tag{D-12}$$

Substituting this and its derivative into Eq. D -9, we get an expression for $\overline{w}(C)$ such that

$$\overline{w}(C) = K_1^2 K_2 K_3\left(1 - K_2 e^{-K_3 C}\right)e^{-K_3 C} \int_0^{2\pi} p(r,\theta)d\theta \tag{D-13}$$



NOTES:

1) Δr IS THE DISTANCE FROM THE PEAK TO THE FIRST ISOPLETH, I.E., $C_p - \Delta C$ (the 0.20 ppm ISOPLETH)
2) PEAK CONCENTRATION OF OZONE IS 0.24 ppm.
3) BACKGROUND CONCENTRATION IS 0.04 ppm.
4) POLLUTANT CLOUD RADIUS IS 13 MILES.

FIGURE D-5. TYPICAL RADIAL CONCENTRATION DISTRIBUTIONS ABOUT THE PEAK. Parameters are chosen to be representative of the Denver metropolitan region.

We now make another key simplifying assumption. We approximate the value of the integral by assuming a uniform radial population density, i.e.,

$$\int_0^{2\pi} p(r,\theta)d\theta = 2\pi D \quad . \tag{D-14}$$

Substituting this into Eq. D-13, we arrive at the final form for $\overline{w}(C)$:

$$\overline{w}(C) = K_0\left(1 - K_2 e^{-K_3 C}\right)e^{-K_3 C} \tag{D-15}$$

where

$$K_0 = 2\pi D\left(\frac{\Delta r}{1 - e^b}\right)^2\left(\frac{b}{\Delta C}\right)\exp\left(\frac{C_P}{\Delta C/b}\right) \tag{D-16}$$

$$K_2 = \exp\left(\frac{C_P}{\Delta c/b}\right) \tag{D-17}$$

$$K_3 = \frac{b}{\Delta c} \tag{D-18}$$

and D is chosen such that the integral of $\overline{w}(C)$ between C_B and C_P equals the total population within the modeled area.

We have made thus far a number of significant assumptions. To test their adequacy, we can select parameter values appropriate for the Denver example, calculate $\overline{w}(C)$, and compare the results against the data shown in Figure D-3. The parameter values selected are shown in Table D-1.

In Figure D-6 we show the population distribution predicted by Eq. D-15. Several observations can be made about its agreement with the test data.

TABLE D-1. SELECTED PARAMETER VALUES IN DENVER TEST CASE

Symbol	Description	Value
C_P	Peak concentration (ozone).	0.24 ppm
C_B	Background concentration.	0.04 ppm
ΔC	Concentration decrement between isopleth lines (N=5 isopleths).	0.04 ppm
b	Exponent by which interisopleth distance grows, selected such that C(r) equals C_B at r=13 miles from the peak (at the approximate urban radius).	0.4
Δr	Radius from peak to the first isopleth (the 0.20 ppm contour).	1 mile
D	Uniform population density chosen such that the integral of $\overline{w}(C)$ between C_P and C_B equals the total population (1.275 million).	2405 persons/ sq. mi.

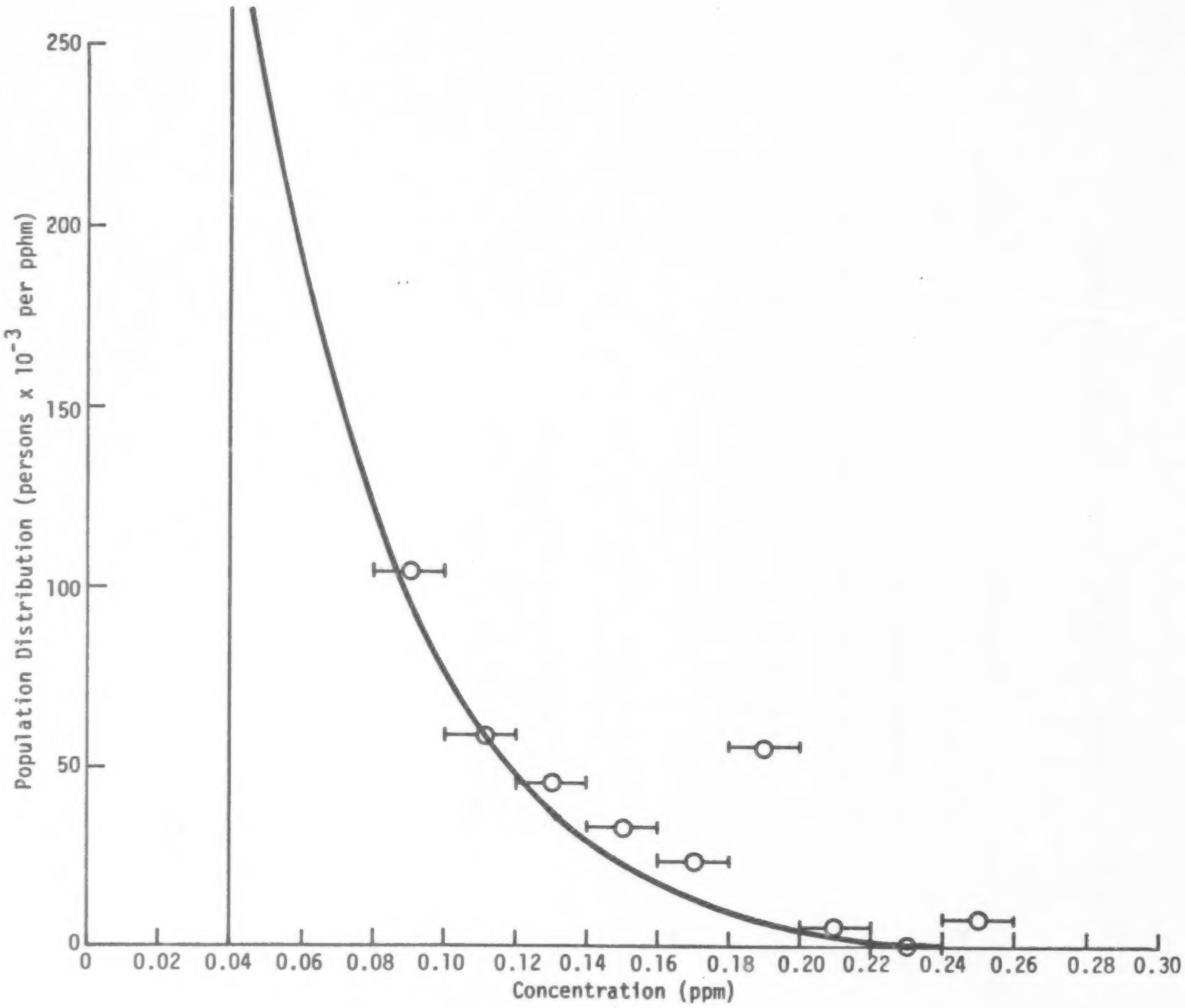


FIGURE D-6. PREDICTED POPULATION DISTRIBUTION AS A FUNCTION OF CONCENTRATION

- Qualitatively, the shapes seem to agree.
- The analytic form of $\overline{w}(C)$ seems to underpredict the distribution of population at higher concentration levels.
- The anomaly occurring in the data at 0.19 ppm remains unaccounted for in the analytic form.

Despite the seeming limitations imposed by our assumptions, however, agreement with the test data seems surprisingly good. It remains to be seen in further investigation (beyond the scope of this study) whether this result is typical or merely fortuitous. We emphasize that results obtained thusfar, while encouraging, should be regarded as preliminary.

In deriving Eq. D-15, we assumed a uniform population distribution. We can estimate qualitatively from our results the change in $\overline{w}(C)$ resulting from variations in this assumption. The shifts expected in $\overline{w}(C)$ for a nonuniform population density are illustrated in Figure D-7. In all cases the integral of $\overline{w}(C)$ is assumed to equal the total regional population.

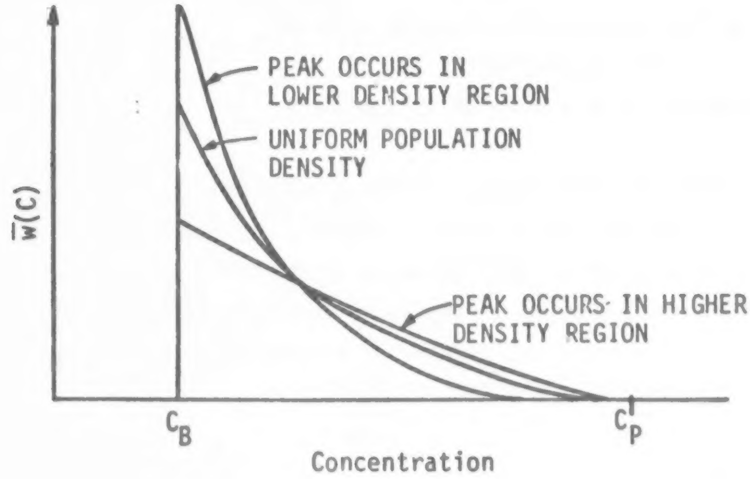


FIGURE D-7. SHIFTS IN $\overline{w}(C)$ CAUSED BY NONUNIFORM POPULATION DISTRIBUTIONS

We now consider other variation of $w(C,t)$ with time. Temporal changes in the function are caused by two principal effects:

- > Evolution of the Concentration Field
  - The peak concentration occurring at a time t, $C_p(t)$, increases during the morning, usually reaches a diurnal peak in the early afternoon, and then decreases slightly by late afternoon.
  - The overall radius of the pollutant cloud--$r(C_B)$--increases up to the time of the peak.
  - As the day progresses near-peak concentrations "spread out," that is, the percentage of the total cloud area having concentrations near the current-hour peak (say, within 20% of it) increases during the day
- > Population Shifts
  - Urban areas have two distinct patterns of population distribution during the day: residential (non-work) and employment (workday). These are separated by two peak-traffic commute periods.
  - A percentage of the population during the day is mobile, traveling from one point to another.

We have assumed here that the total impact of these effects can be approximated by a separable weighting function, $f_w(t)$, applied to the function $\overline{w}(C)$. The extent to which this is valid needs to be verified by additional investigation. Yet, as a first approximation it has some plausibility, and it allows us to proceed to an analytic result for model performance standards--our principal objective.

Health Effects Function

Health effects resulting from exposure to polluted air manifest themselves in many ways, each varying in the symptom it produces and the seriousness of its impact. Among such effects are the following: bronchial irritation, reduced lung function, enzyme damage, eye irritation, dizziness, and coughing. Some of these manifest themselves as noticeable but low-level discomfort; others produce more serious impact such as aggravation of respiratory illness. Equating each effect on an absolute scale and relating their aggregate weighted impact directly to ambient pollutant levels, however, is a formidable task. Efforts at doing so have been subject to uncertainty and controversy. To overcome these difficulties, we resort to several conceptual simplifications. Rather than differentiating between individual health effects, we collapse them together into a single function, whose "seriousness" is dependent on concentration level, C, and duration of exposure, Δt. We represent this by the following

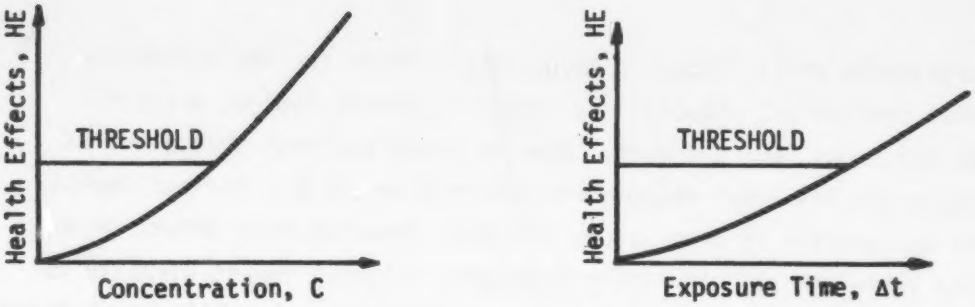
$$HE = HE(C,\Delta t) \quad . \tag{D-19}$$

We now make an intuitive appeal. While we may not know the value of HE in an absolute sense, we observe that its value increases, that is, the HE gets "worse," as concentration levels rise and the duration of exposure increases. Further, because health effects at higher concentrations and durations are more serious, we expect HE to grow faster than linearly with increasing C (and probably Δt). We also can expect HE to exist even at very low values of C, though these effects may be small, perhaps below the threshold of human perception. Qualitatively, the shape of HE might look as shown in Figure D-8.

Based on the reasons noted above, we can make a useful approximation. We assume that HE is separable, one part dependent on C and Δt, and that it can be described by the following simple relation:

$$HE(C,\Delta t) = AC^{\gamma}f_{HE}(\Delta t) \quad , \tag{D-20}$$

where A is a scaling constant (whose value we need not know, as we shall observe later); γ is a "shaping" parameter whose value is likely to be



(a) Variation With Concentration (b) Variation With Exposure Time

FIGURE D-8. EXPECTED SHAPE OF HEALTH EFFECTS FUNCTION

greater than one, i.e., linear; and $f_{HE}(\Delta t)$ is a weighting function dependent solely on exposure time.

c. Analytic Solution of the Cumulative Health Effects Functional

Having now specified analytic forms for the population distribution function, $w(C,t)$, and the health effects function, $HE(C,\Delta t)$, we may proceed to evaluate the area-integrated cumulative health effects functional, ϕ, as it was defined in Eq. D -6. We may rewrite ϕ as follows:

$$\phi(C_p\ \Delta t) = \int_{t_1}^{t_2} f_w(t) f_{HE}(t - t_1)dt \int_{C_B}^{C_p} \overline{w}(C) A C^{\gamma} dC$$

$$= F(\Delta t)\psi(C_p) \quad , \qquad \text{(D-21)}$$

where C_p is the peak concentration experienced during the day.

Using relations developed previously, we may evaluate ψ. Its value is

$$\psi(C_p) = A \int_{C_B}^{C_P} \overline{w}(C) C^{\gamma} dC$$

$$= A \int_{C_B}^{C_P} \left[K_0 \left(1 - K_2 e^{-K_3 C}\right) e^{-K_3 C} \right] C^{\gamma} dC \qquad \text{(D-22)}$$

Though no completely general solution exists to this equation, the integral may be evaluated in closed-form for each integer value of γ, the health effects function shaping parameter. A point-wise analytic solution to Eq. D -22 thus exists.

d. Calculation of Minimum Allowable Predicted Peak

As noted in Eq. D-4, the model performance standard could be specified in terms of a minimum allowable ratio of the "predicted" to "measured" values of ϕ. If that ratio is r, then the following relationship would exist at the minimum acceptable level of model performance:

$$r = \frac{\phi\ (C_{p_P}\ ,\ \Delta t)}{\phi\ (C_{p_m}\ ,\ \Delta t)}$$

$$= \frac{F\ (\Delta t)\ \psi(C_{p_P})}{F\ (\Delta t)\ \psi(C_{p_m})}$$

$$= \frac{\psi\ (C_{p_P})}{\psi\ (C_{p_m})} \qquad \text{(D-23)}$$

where C_{P_p} is the predicted peak concentration and C_{P_m} is the measured peak value. By writing the standard in this form, an important simplification results: Two parameters, being constant, appear outside the integrals in the numerator and denominator of Eq. D-23. Since their values in both are equal, they cancel. By this means, we eliminate the need for "knowing" the health effects function scaling coefficient, A, and the population distribution scaling constant, K_0. With the rationale we present here, uncertainty associated with both, while appreciable, thus does not affect the setting of performance standards.

We can invert Eq. D-23 to solve for the minimum allowable ratio of predicted to measured peak concentration value. We do so for the Denver example discussed earlier, presenting the results in Figure D-9. We show results for several representative values of γ and r. If health effects varied linearly with concentration and r equaled 0.90, for instance, any predicted peak would be acceptably higher than 64 percent of the measured peak value. Similarly, if health effects were a cubic function of concentration and r=0.90, the predicted peak would have to exceed 80 percent of the measured value.

Several decisions must be made in determining a final value for a performance standard based upon this health effects rationale: A minimum acceptable value must be chosen for r, the ratio of predicted to measured area-integrated cumulative health effects; and a judgment must be made about the maximum likely value of γ, the exponent of concentration in the health effects function. Possible values for use might be r and γ of 0.90 and 3 or 4 respectively. For reference, we note that for $\gamma = 10$, the minimum allowable ratio of predicted to measured peak is 94 percent.

e. The Health Effects Rationale: A Summary

A model performance standard based upon pollutant health effects has intuitive appeal. For this reason the rationale presented in this

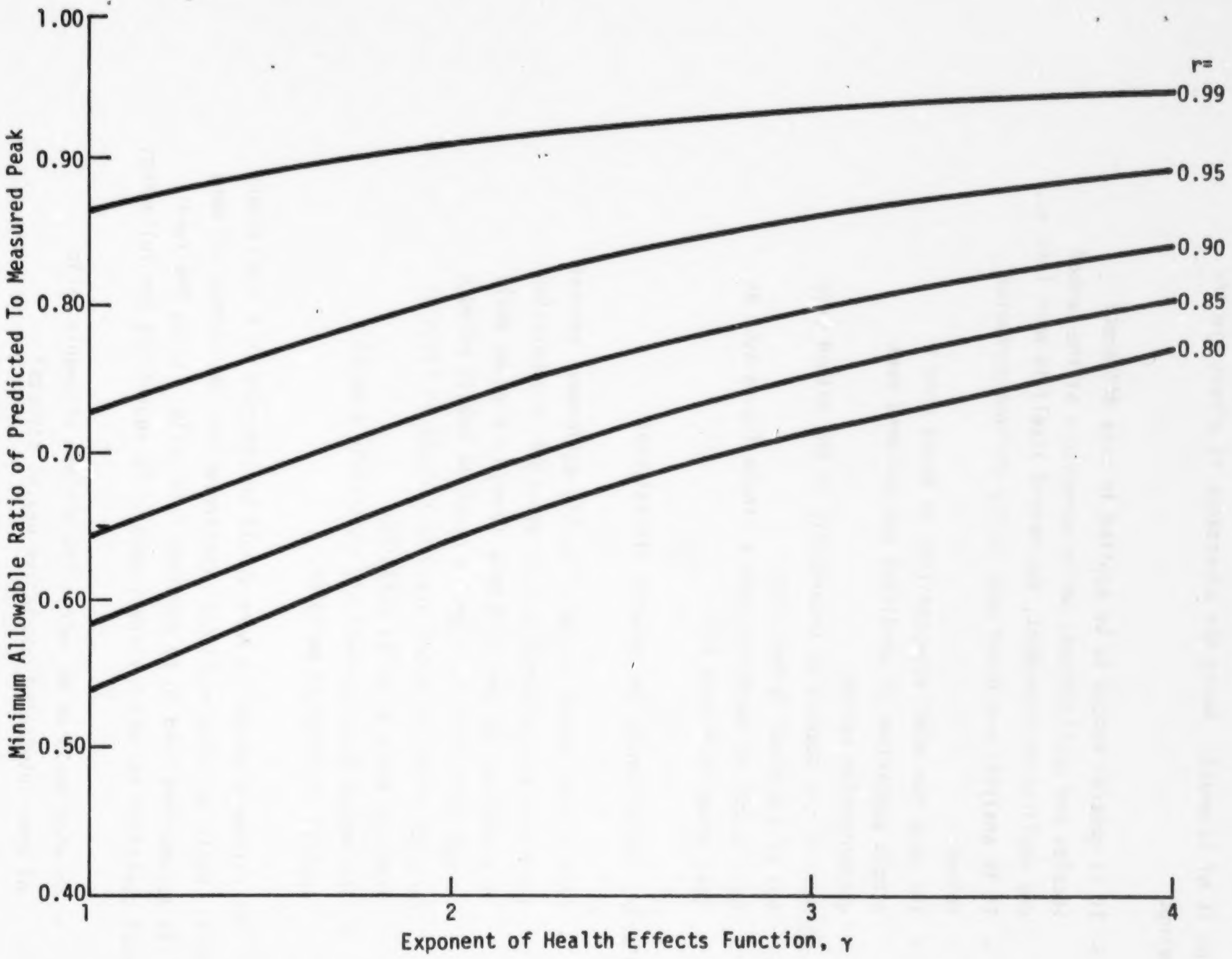


FIGURE D-9. MINIMUM ALLOWABLE RATIO OF PREDICTED TO MEASURED PEAK CONCENTRATION VALUE section is of interest. Among the advantages it offers are the following:

- It is general enough to be applied in many different locales and applications; while parameters of the method are application-dependent, the method itself is much less so.
- It is analytic and based upon easily derived parameter values.
- The test for model acceptability is based upon a simple comparison of predicted and measured peak concentration values.
- Many of the sources of uncertainty in the method drop out of its final formulation.
- Results can be condensed into a single figure such as that shown in Figure D-9.

Similarly, the rationale has several limitations:

- Only a lower bound on the allowable difference between predicted and measured peak is provided; a prediction in excess of the measured peak (even by a great deal) is not sufficient to reject a model on health effects grounds since the model predicts effects at least as great as those actually existing.
- The method does not evaluate explicitly a model's spatial or temporal behavior.

The rationale presented here should be regarded as a preliminary method. While meriting additional consideration, the method and many of its assumptions need to be examined critically. Among the fundamental questions for which answers need to be sought are the following:

- On what basis do we select the minimum allowable ratio of area-integrated cumulative health effects?

> What value of health effects exponent is most appropriate?

> Does the population distribution, $\bar{w}$ (C), always reproduce the data as well as indicated in Figure D-6? Does it need to?

> Is w(C,t) really a separable function, as assumed? What about HE(C,Δt)?

> Are health effects really related to peak concentration and exposure time in the fashion assumed here? What about those who work in environmentally controlled buildings and may thus be isolated from full exposure to ambient concentration levels?

We feel the rationale presented here has a number of advantages. We also feel it requires a careful review and some additional examination, particularly as regards the questions noted above.

2. Control Level Uncertainty Rationale

In order to reduce peak ambient concentrations in an airshed from a particular level to one at or below the NAAQS, reduction of emissions into that airshed is required. The degree of that reduction, however, is dependent on the amount by which the current peak level exceeds the standard. Uncertainty in our knowledge of the current peak concentration (due either to measurement or modeling limitations) translates into corresponding uncertainty in the amount of emissions control we must require. This direct relationship, though generally a highly nonlinear one, forms the basis for another rationale for setting model performance standards. Its guiding principle is as follows: Uncertainty in the percentage of emissions control required (PCR) must be kept to within certain allowable bounds.

In this section we discuss this Control Level Uncertainty (CLU) rationale. We first indicate for a specific pollutant (ozone) how one may proceed from PCR bounds to equivalent allowable tolerances on the difference between the predicted and measured peak concentration. We then present one means whereby the PCR bounds can be determined from the economies of pollution control costs. Several benefits derive from use of the CLU rationale, among which are the following:

- > It makes explicit the relationship between model performance limits and the maximum acceptable level of uncertainty in estimates of regional emissions control.
- > It provides a structure whereby model performance limits also can be related to equivalent uncertainty bounds on the total regional cost of pollution control equipment.

The rationale presented here is a useful complement to the Health Effects (HC) rationale presented earlier. We noted in discussion of that rationale that it could not provide an upper bound on the maximum allowable difference between predicted and observed peak concentration levels. It merely required that the predicted peak be greater than a fraction (near unity) of the measured peak, i.e., $C_{p_p} \geq \beta C_{p_m}$ where β is near unity (e.g., 0.9). Were C_{p_p} to be larger than C_{p_m}, no health effect penalty would be incurred by designing a control strategy based upon C_{p_p}. Rather, the principal penalty would be an economic one: The cost of control would be greater than that actually required. It is in setting the upper bound on the allowable value of $C_{p_p} - C_{p_m}$ that the CLU rationale has its greatest value, since it addresses directly the cost of control.

We can generalize this point as follows: The greatest cost of underprediction of the peak concentration lies in the underestimation of health impact, while the greatest consequence of overprediction is the extra economic cost associated with unnecessarily imposed control. Health Effects and CLU, then, are compatible rationales. If the predicted peak is required to statisfy $K_1 \leq C_{p_p} - C_{p_m} \leq K_2$, then it seems reasonable that K_2 be selected based upon the CLU rationale with K_1 chosen to be the lesser of the values determined by the HE and CLU rationales.

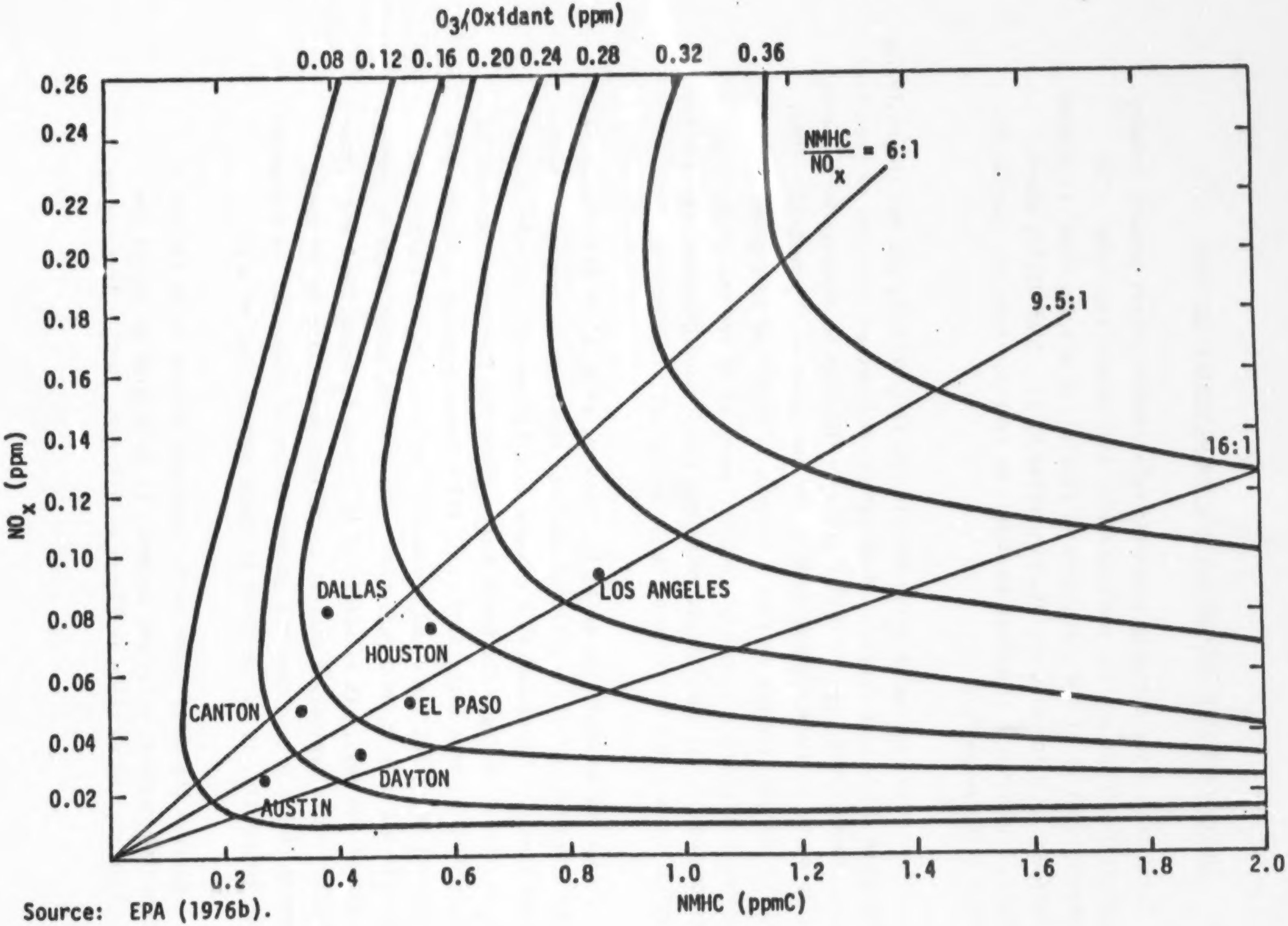
a. The Relationship Between CLU and the Concentration Peak

In most cases a highly nonlinear relationship exists between primary emissions and the ambient concentrations that result from them. The dynamic behavior of the atmosphere is complex, as are the chemical changes undergone by dispersing pollutants carried by it. Simplifying assumptions, however, can sometimes be made. We consider here one example in which this can be done.

For urban regions in which certain specific criteria are met (Hayes,1977), the ozone production resulting from various non-methane mixtures of precursor hydrocarbons (NMHC) and oxides of nitrogen (NO_x) can be represented by means of an ozone isopleth diagram such as the one shown in Figure D-10. (EPA, 1976). Whether the use of such a diagram is justified in a given region depends heavily on a number of factors, among which are the prevailing meteorology, solar insolation, emissions type/timing/geometry, terrain type/complexity, and the presence of large upwind pollutant sources.

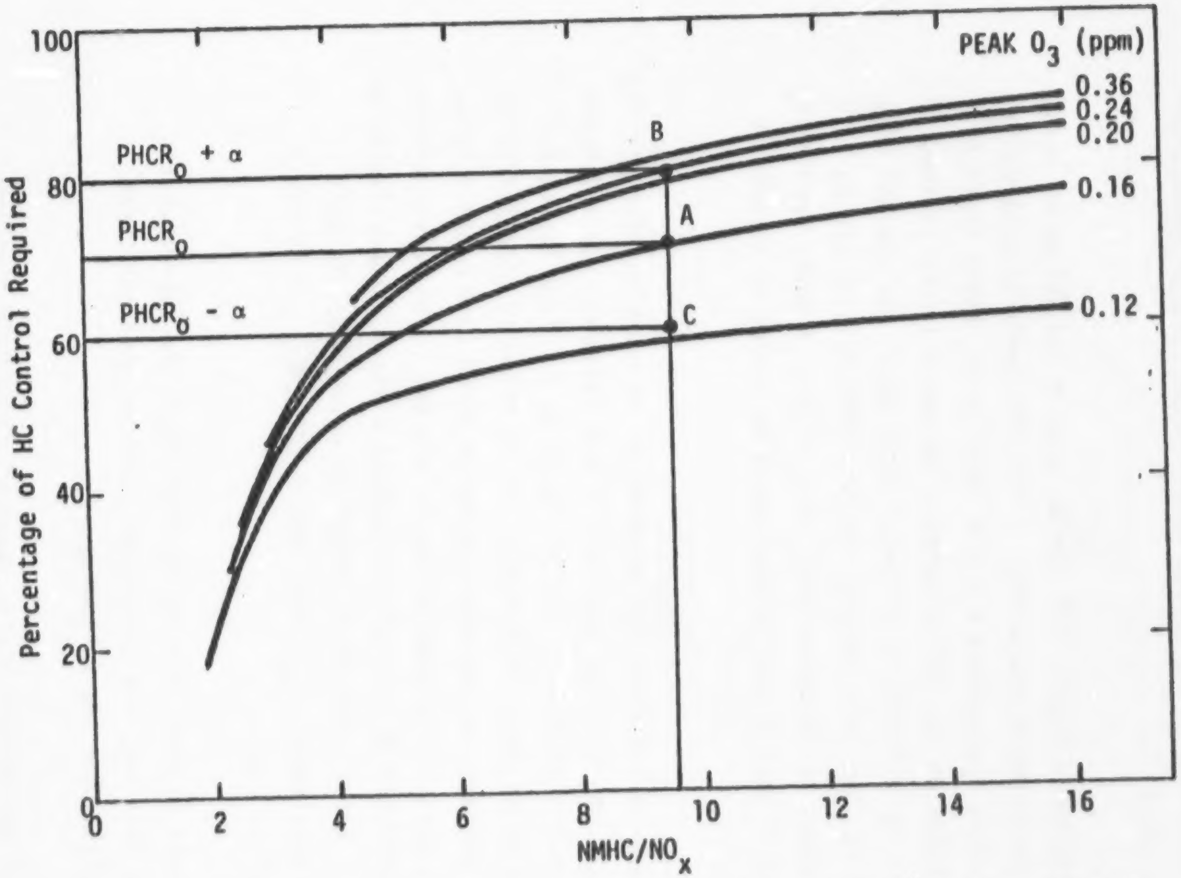
If a region meets the criteria, however, an isopleth diagram may be used as an approximation relating regional emissions to consequent peak ozone levels. The region-wide cutback in emissions of precursor HC and NO_x necessary to reach the NAAQS from a given starting point can then be calculated, given a background ozone value (usually about 0.04 ppm) and a control mix (NMHC versus NO_x cutback). Usually, in urban areas the emphasis has been on NMHC reduction. The starting point often is defined in one of two ways: It is specified by a peak O_3 measurement and either a NMHC/NO_x ratio typical of ambient conditions prevailing in the early morning (6-9 a.m.) or specific concentrations of either of the precursors. Most frequently, it is the first of these methods that is used.

Because the chief value of the isopleth diagram is in its use in estimating regional emissions cutback, it is helpful to replot the isopleth diagram as shown in Figure D-11 (Hayes, 1977). In doing so,



Source: EPA (1976b).

FIGURE D-10. PROTOTYPICAL ISOPLETH DIAGRAM



Note: No change in NO_x level and no O_3 background concentration were assumed.

FIGURE D-11. THE ISOPLETH DIAGRAM REPLOTTED percentage control required (PCR) can be highlighted explicitly. While in principle any mix of NMHC and NO_x control could be considered, the example shown assumes that only HC control is employed. That is, percentage control reduction (PCR) is equivalent to percentage hydrocarbon control required (PHCR).

The PHCR diagram in Figure D-11 may be used in the following way to deduce model performance standards. First, the measured peak ozone concentration and the appropriate 6-9 a.m. NMHC to NO_x ratio together define a unique point on the PHCR diagram. The nominal PHCR is thus identified. Then, by defining an allowable band about the nominal PHCR (say $\pm \alpha$ where α is some small value), we can identify directly an equivalent band about the measured peak ozone value. A model predicting an ozone peak within that allowable band would be judged as acceptable under this rationale.

We can illustrate the technique by means of an example. Suppose the measured peak ozone was 0.16 ppm and the 6-9 a.m. $NMHC/NO_x$ was estimated to be 9.5. This point is denoted on the figure as A. From Figure D-11, we see that the PHCR is about 70 percent. If we allow an uncertainty in the PHCR of ± 10 percent, we see that the value based upon model predictions of the peak must lie between 60 and 80 percent. The corresponding values of peak ozone are determined from points C and B, respectively, on the PHCR diagram. For a model to be judged as acceptable, it must predict an ozone peak value, C_{p_p}, such that $0.122 \leq C_{p_p} \leq 0.24$ ppm or $76 \leq C_{p_p}/C_{p_m} \leq 150$ percent.

Several general observations may be made about the above results, though we caution that they are particular to ozone as a pollutant. Among the observations are the following:

- Because of the characteristic shape of ozone PHCR diagrams, the upper value of the allowable tolerance band is less restrictive then the lower one. This is illustrated clearly in the example.

> The allowable band for C_{p_p} is always bounded on the upper and lower side (as contrasted with the HE rationale which calculates only a lower bound).
> In those cities for which use of the ozone isopleth shown in Figure D-11 is appropriate and where the 6-9 a.m. NMHC/NO_x is greater than about 5 or 6, the width of the allowable band for C_{p_p} is not strongly sensitive to the value of NMHC/NO_x.

b. The Relationship Between CLU and Control Cost

While the allowable uncertainty in control level ($\pm \alpha$ in the above example) may be set in many ways, we examine here one important means to do so: the explicit use of regional pollution control costs, if these can be specified unambigously. We might, for instance, choose as our guiding principle the following: The uncertainty in the total cost of regional pollution control should not be greater than a certain value δ. We may restate this in terms of model performance. The level of control deriving from the predicted peak, C_{p_p}, should not differ in cost by more than a certain amount from that level determined based upon the measured peak, C_{p_m}.

To proceed we must define the total regional cost of pollution control, TC. Depending on the level of control required, alternative regional control strategies can be designed. The cost of each generally can be specified, at least in approximate terms. By plotting the cost of a series of "preferred" strategies against the level of control they achieve, TC can be determined, as shown in Figure D-12.

Several aspects of the TC curve should be noted. While TC is zero for a PCR of zero, any non-zero value of PCR has associated with it a minimum, non-zero cost. Thus, the TC curve really "begins" with a step function at PCR = 0. TC rises quickly at first as many fixed costs of control are incurred. The cost then increases more slowly as fixed costs are spread over greater values of PCR. Finally, at high levels of PCR, each additional amount of control becomes more difficult (and more expensive) to achieve. The TC function, consequently, rises rapidly.

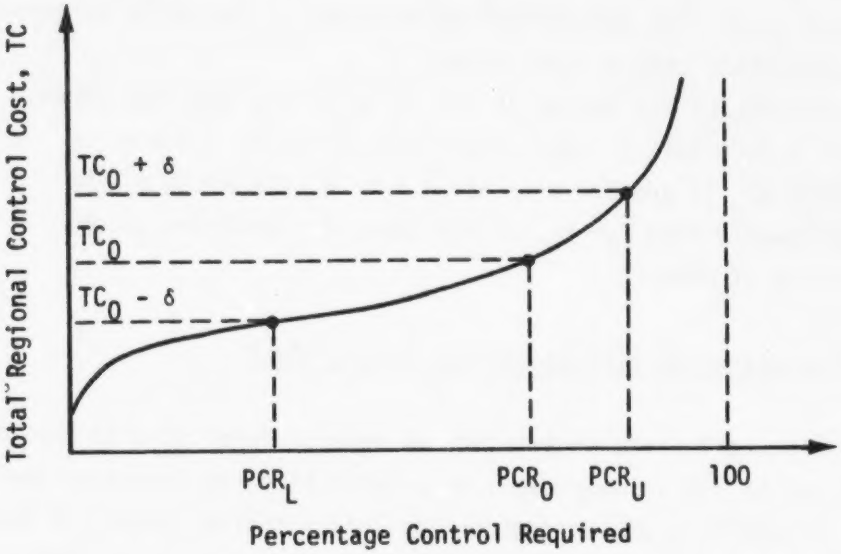


FIGURE D-12. TOTAL REGIONAL CONTROL COST AS A FUNCTION OF THE LEVEL OF CONTROL REQUIRED

Once the total cost function has been defined, the allowable band for the predicted ozone peak can be found in the following way:

> Step 1. The nominal control level PCR$_0$ can be determined using a diagram such as that in Figure D-10. With all-NMHC control as considered in deriving that figure, PCR$_0$ is identical to PHCR$_0$,

> Step 2. The nominal control cost, TC$_0$, can be found using a TC diagram similar to the one in Figure D-11.

> Step 3. The maximum and minimum allowable TC values then can be calculated and the corresponding bounds on PCR determined.

> Step 4. Using the PHCR diagram once again, the allowable bounds on predicted peak ozone can be found by employing the PCR bounds found in Step 3.

The above procedure is a straightforward one creating a structure in which control cost uncertainty can be considered explicitly. The example presented, however, is appropriate only for considering ozone in those regions having ambient conditions simple enough to be represented by an isopleth diagram. Extension of the procedure to other pollutants and into regions of greater atmospheric complexity requires that additional research be conducted beyond the scope of the current effort.

3. Guaranteed Compliance Rationale

As formulated in the federal regulations, the NAAQS are explicit, with maximum pollutant levels specified that must not be exceeded with greater than a certain frequency. Peak one-hour concentrations of ozone, for instance, must not exceed 0.08 ppm more often than once per year. With the standards written in such an absolute fashion, it may be argued that little room exists for uncertainty about achieving compliance. Under such circumstances, a model's performance should be constrained to "guarantee" that its use will not lead to underestimating the degree of emissions control required.

Model behavior can affect significantly the likelihood of meeting the NAAQS. In those regions currently in noncompliance, the effectiveness of candidate control strategies can be assessed only by means of model predictions of the peak concentrations resulting from each. If a model systematically underpredicts the peak value for concentrations near the NAAQS, the adequacy of controls might be overestimated. Similarly, if the model overpredicts the peak, controls designed using it might be excessive.

a. Description of the GC Rationale

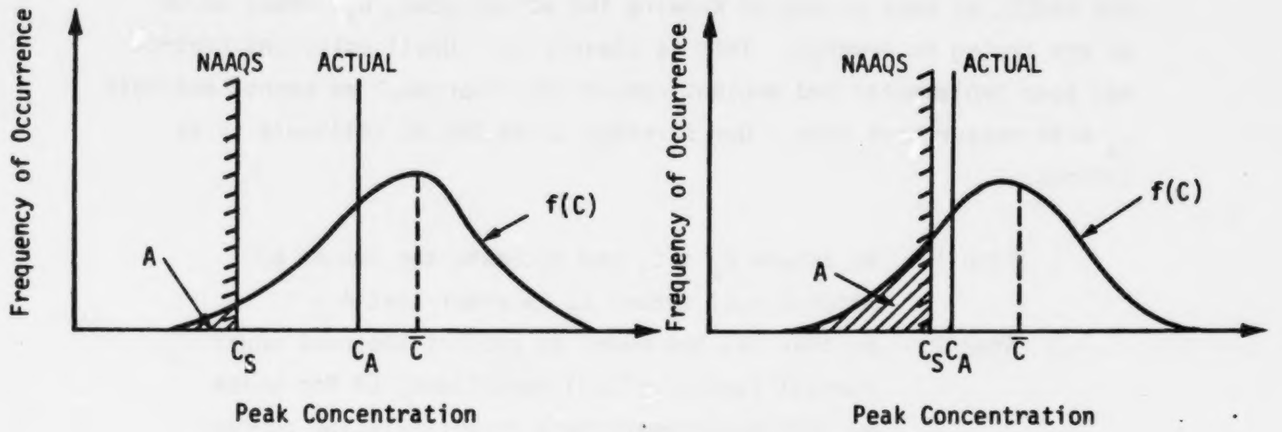
With the above in mind, we examine the Guaranteed Compliance (GC) rationale for setting model performance standards. We state its guiding principle as follows: Compliance with the NAAQS must be "guaranteed," with all model uncertainty on the conservative side even if it means introducing a systematic bias into model predictions. The term "guaranteed" should be taken here in a limited sense. We intend it to mean that "the probability is very small" that a model will predict a peak value less than the standard when its actual value is greater.

We illustrate this principle using the diagrams in Figures D-13 and 14. In these figures we illustrate two models, one "conservative" (Figure D-13) and the other "nonconservative" (Figure D-14). For each, we show two cases: an actual peak concentration, C_A, higher than the NAAQS, C_S, and one near the standard. We represent the probability density function of the model as f(C) and the expected value of the predicted peak as $\overline{C}$. Two types of uncertainty affect a model's performance. The first includes error in model inputs and uncertainty in the values of the model parameters themselves. These affect the shape of f(C). Uncertainty of the second type is due to the inability of the model formulation to represent reality fully. The difference between the expected model prediction, $\overline{C}$, and the actual value, C_A, of the peak concentration is a measure of the effect of formulation errors. As we define it here, a "conservative" model is one for which the value of $\overline{C}$ exceeds C_A, while for a "nonconservative" model the reverse is true. In both figures, the shaded area A represents the probability that the model will predict a peak concentration less than the standard at the same time the actual value is greater.

With the GC rationale, we want to insure that A remains acceptably small. In mathematical terms, we insist that

$$A = \int_{-\infty}^{C_S} f(C)dC \leq \xi \quad , \qquad \text{(D-25)}$$

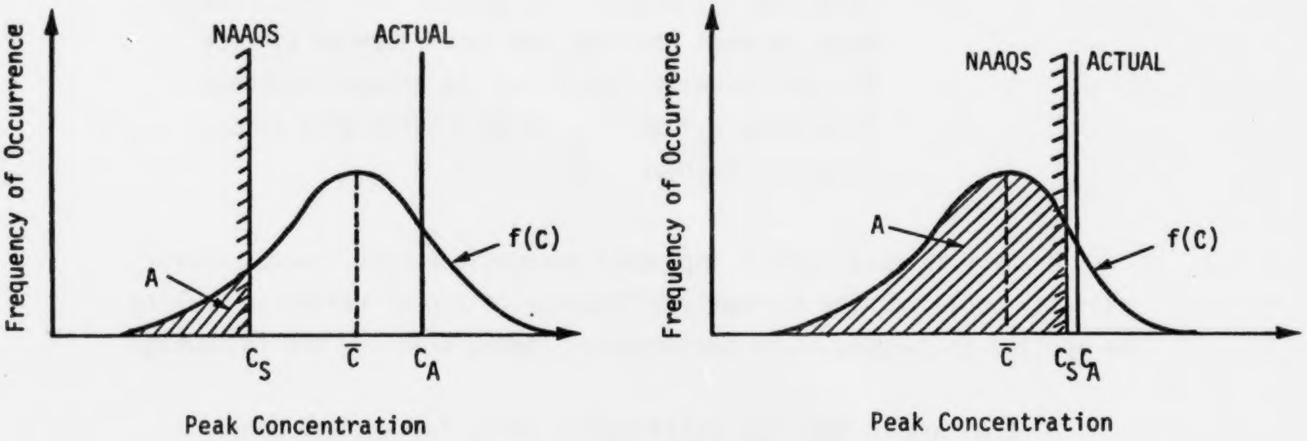
where ξ is some suitably small number. From the figures we see that A can be kept small only if $\overline{C}$ exceeds C_A. Under the requirements of the GC rationale, only a model having these characteristics would be judged acceptable.



(a) Peak Concentration Higher than the NAAQS

(b) Peak Concentration Near the NAAQS

FIGURE D-13. UNCERTAINTY DISTRIBUTION FOR A CONSERVATIVE MODEL



(a) Peak Concentration Higher than the NAAQS

(b) Peak Concentration Near the NAAQS

FIGURE D-14. UNCERTAINTY DISTRIBUTION FOR A NONCONSERVATIVE MODEL

A practical consideration now becomes important. For peaks near the NAAQS, we have no way of knowing the actual peak, C_A, whose value we are trying to predict. This is clearly so. Until emissions control has been implemented and ambient conditions "improve," we cannot estimate C_A with measurement data. Our strategy using the GC rationale is as follows:

> Step 1. We assume $C_A = C_S$ and estimate the amount by which $\overline{C}$ must exceed C_A in order that $A \leq \xi$.

> Step 2. We then use the model to predict the peak under current (uncontrolled) conditions, $\overline{C}^*$ for which we have measurement data to estimate the current peak, C_A^*.

> Step 3. To judge acceptability, we require the model prediction, $\overline{C}^*$, to exceed C_A^* by as much as $\overline{C}$ exceeded C_A when $C_A = C_S$. Actually, this is a bit more complicated. Since C_A^* is based upon measurements, it is subject to instrumentation error. We know C_A^* only in terms of a measured value and its probability density function. Therefore, we must consider the comparison of C_A^* and $\overline{C}^*$ statistically, requiring the probability that $\overline{C}^*$ exceeds C_A^* by $\overline{C}-C_A$ to be greater than some large value (near 1.0).

We have invoked several important assumptions here, whose general validity would require further verification if the GC rationale were to be applied in judging model performance. Among them are the following:

> $\overline{C}$ maintains the same relationship to C_A for ambient conditions ranging from current ones to those characterizing compliance with the NAAQS.

> The probability density function, f(C), is known or can be determined, as can $\overline{C}$.

> Instrumentation uncertainty can be characterized, allowing Step 3 to be accomplished.

There are several difficulties associated with the GC rationale approach, however, some of which are conceptual and some practical. Among the most important of the conceptual difficulties is the introduction of a conservative bias into model predictions. By insisting that the model "overpredict" peak concentrations, almost certainly we will select abatement strategies requiring more control than needed. Difficulties of the practical kind also can be significant. For most models, determination of f(C) is a difficult (and usually impractical) process. The uncertainty in predicting the peak is partially due to uncertainty in the data input to the model. Since the model results are related to inputs only in a complex and nonlinear way, estimating the output uncertainty distribution in terms of the input error distributions seldom can be done directly. While a Monte Carlo-type of analysis in principle can be conducted, the number of model runs required and the amount of computing resources consumed are so considerable as to render such an analysis impractical.

b. A Possible Simplification

Short of doing a Monte Carlo analysis, is there anything useful that can be determined? In certain simple circumstances, there is. We may infer, when appropriate, some limited information about f(C), $\overline{C}$ and C_A. To do so, we first recall the modified form of Tchebycheff's inequality,

$$P\{|\underline{\chi} - \eta| \geq k\sigma\} \leq \frac{4}{9k^2} \quad , \qquad \text{(D-26)}$$

where P is the probability that $-k\sigma \geq \underline{\chi} - \eta$ and $k\sigma \leq \underline{\chi} - \eta$, $\underline{\chi}$ is a random variable, η is its expected value, and σ is its standard deviation. This relationship holds for all probability distributions. We can adapt it to the present problem by rewriting it in the following way:

$$P\left\{C - \overline{C} \leq C_s - \overline{C}\right\} \leq \frac{1}{2} \cdot \frac{4}{9}\Bigg/\left(\frac{\overline{C} - C_s}{\sigma_c}\right)^2$$

$$\leq \frac{2}{9}\left(\frac{\sigma_c}{\overline{C} - C_s}\right)^2 \quad , \qquad \text{(D-27)}$$

where C is a random variable whose value is the peak concentration predicted by the model, $\overline{C}$ is its expected value, and σ_c is its standard deviation. C_s is the standard (NAAQS).

The relation in Eq. D-27 is a useful one. The area A in Figures D-13 and 14 represents the same probability as that on the left hand side of Eq. D-27. Using Eq. D-25, we may now write

$$A \leq \xi \leq \frac{2}{9}\left(\frac{\sigma_c}{\overline{C} - C_s}\right)^2 \quad , \qquad \text{(D-28)}$$

where ξ is the maximum allowable value of A. From this, we may infer the minimum allowable value of $\sigma_c/(C_s - \overline{C})$. Its value is

$$\left(\frac{\sigma_c}{\overline{C} - C_s}\right)_{min} = \sqrt{\frac{9}{2}\xi} \quad . \qquad \text{(D-29)}$$

Still, we need an independent approximation of σ_c in order to solve Eq. D-29 for the minimum value of $\overline{C} - C_s$. To do so, we estimate the maximum value σ_c is likely to assume, that is, the σ_c^* such that

$$\sigma_c \leq \sigma_c^* \quad . \tag{D-30}$$

If we then use σ_c^* in Eq. D-29, we can determine $(\overline{C} - C_s)_{min}$.

Suppose we represent model behavior with a system response function, ϕ, that tranforms model inputs into the model-predicted concentration peak, i.e.,

$$C = \phi(\underline{\varepsilon}) \quad , \tag{D-31}$$

where C is the predicted peak, an $\underline{\varepsilon}$ is the vector of model inputs. Suppose further that we know the probability distributions of each of the input errors, and that we can identify their one-sigma variations, σ_{ε_i}. If so, we can determine the maximum change in the predicted peak that would occur if all error sources varied simultaneously by a standard deviation from their nominal values. We note that increases in some inputs lower C and others raise it. Thus, to bound the value of ΔC, we consider the root-mean-square of the changes in C as each input is varied separately. This maximum ΔC can be written as

$$\Delta C = \sqrt{\sum_{i=1}^{N} \left(\phi|_{\varepsilon_i + \sigma_{\varepsilon_i}} - \phi|_{\varepsilon_i}\right)^2} \tag{D-32}$$

where each ε_i $(1 \leq i \leq N)$ is varied separately and the corresponding change in peak concentration is represented by the quantity in the brackets. If we assume that ΔC is a suitable estimate of σ_c^*, we can write (using Eq. D-29)

$$(\overline{C} - C_s)_{min} = \frac{\Delta C}{\sqrt{\frac{9}{2}\xi}} \quad , \tag{D-33}$$

which provides an indication of the amount of "overprediction" the model must provide.

We now present an example. Suppose we consider a simple Gaussian model (no reflection, continuously emitting source), whose only source of error is the wind speed, U. We assume the following: $\sigma_U = 0.5$ m/sec, $U = 2$ m/sec, and $C_s = 35$ ppm (the one-hour federal standard for CO). Using Eq. D-32, we determine that $\Delta C = 7$ ppm. Then, using Eq. D-33 and assuming that $\xi = .05$, we estimate that $(\overline{C} - C_s)_{min} = 14.7$ ppm. Using the GS rationale, we would require when modeling current ambient conditions that the model overpredict the peak by this same amount (assuming that there was no error associated with the measurement).

c. The GC Rationale: An Assessment

We have included the GC rationale in our discussion primarily for the sake of completeness. While the guiding principle underlying it--"guaranteeing" that an adequate abatement strategy will be designed--has its virtues, the method as conceived here has significant problems associated with its use. It is cumbersome and impractical, except in the most limited of circumstances. Also, it may be excessively conservative, introducing a systematic bias into model evaluation.

Unless the major problems noted here can be solved somehow, the other rationales considered in this chapter appear to have greater promise. We do not recommend that this rationale be pursued extensively in any additional work.

4. Pragmatic/Historic Rationale

Experience is growing in the use of air quality simulation models. They have been applied to a variety of problems in a number of different situations. As an familiarity grows with both their capabilities and limitations, we become more able to foresee their behavior in new applications. Taking advantage of our growing expertise, we may find it reasonable to set performance standards for models based upon the following principle: In each new application a model must perform at least as well as the "best" previous performance of a model in its generic class in a similar application.

This approach is a pragmatic one, forced upon us by some very practical considerations: our limited ability to derive theoretically justifiable values for the standards and the number of different measures required to characterize fully model performance. Five major problem areas exist in characterizing the agreement of model predictions with field observations. The model may be judged on its ability to predict the concentration peak, to avoid systematic bias, to limit absolute error, to maintain spatial alignment, and to reproduce temporal behavior of concentrations. To assess a model's performance in these five areas, we recommended earlier in this chapter the use of a number of different performance measures. Our chief difficulty is as follows: There are as yet few theoretical means to assign appropriate values for these measures. We have identified in this report several promising candidates for judging the prediction of peak concentrations. Additional work is required, however, to determine appropriate standards for many of the other measures.

While such additional work is proceeding, what must we do? Many issues of great practical interest are pending, each of which requires the evaluation of model performance. Revisions to State Implementation Plans, for instance, must be reviewed. Model performance studies now being conducted by the EPA must continue.

We recommend that the Pragmatic/Historic rationale be used to set acceptable bounds for performance measures for which no other better method exists. As research provides greater insight into "better" rationales, we recommend appropriate updates to the standards.

To employ this rationale the following steps might be followed:

- > Step 1. The proposed application is categorized, identifying the group of previous studies with which its performance must be compared. The criteria by which this might be done could include pollutant type, prevailing meteorology, source geometry, and terrain irregularity.
- > Step 2. Performance measures appropriate to the applications category are calculated.
- > Step 3. Calculated values are compared with the "best" values previously attained in a similar application.

For the Pragmatic/Historic rationale to be of use, the EPA would have to accomplish the following steps. A scheme for classifying applications into "similar" categories needs to be developed. Then, data on previous modeling efforts needs to be assembled and appropriate performance measure values calculated. Finally, a mechanism for updating the "performance data base" needs to be established. Such a mechanism would require the EPA to assume a custodial role over the data base, amending it as results of new modeling studies become available.

REFERENCES

Ames, J., et al. (1978), "The User's Manual for the SAI Airshed Model," EM78-89, Systems Applications, Incorporated, San Rafael, California.

Anderson, G. E. (1978), private communication.

Anderson, G. E., et al. (1977), "Air Quality in the Denver Metropolitan Region 1974-2000," EF77-22, Systems Applications, Incorporated, San Rafael, Calfiornia.

Argonne (1977), "Report to the U.S. EPA of the Specialists' Conference on the EPA Modeling Guideline," 22-24 February 1977, Argonne National Laboratory, Argonne, Illinois.

Burton, C. S., et al. (1976), "Oxidant/Ozone Ambient Measurement Methods," EF76-111R, Systems Applications, Incorporated, San Rafael, California.

Calder, K. L. (1974), "Miscellaneous Questions Relating to the Use of Air Quality Simulation Models," Proc. of the Fifth Meeting of the Expert Panel on Air Pollution Modeling, Chapter 6, NATO/CCMS.

Code of Federal Regulations [CFR] (1975), Title 40 (Office of the Federal Register, U.S. Government Printing Office, Washington, D.C.).

EPA (1977), "Uses, Limitations and Technical Bases of Procedures for Quantifying Relationships Between Photochemical Oxidants and Precursors," EPA-450/2-77-021a, Office of Air Quality Planning and Standards, Environmental Protection Agency, Research Triangle Park, North Carolina.

__________ (1978a), "Workbook for the Comparison of Air Quality Models," EPA-450/2-78-028a,b, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

__________ (1978b), "Guidelines on Air Quality Models," EPA 450/2-78-027, Office of Air Quality Planning and Standards, Environmental Protection Agency, Research Triangle Park, North Carolina.

Johnson, W. B. (1972), "Validation of Air Quality Simulation Models," Proc. of the Third Meeting of the Expert Panel on Air Pollution Modeling, Chapter VI, NATO/CCMS.

Liu, M. K., and D. R. Durran (1977), "The Development of a Regional Air Pollution Model and Its Application to the Northern Great Plains," EPA-908/1-77-001, Office of Energy Activities, U.S. Environmental Protection Agency, Denver, Colorado.

Rosen, L. C. (1977), "A Review of Air Quality Modeling Techniques," UCID-17382, Lawrence Livermore Laboratory, Livermore, California.

Roth, P. M., moderator (1977), "Report of the Validation and Calibration Group (II-5)," in "Report to the U.S. EPA of the Specialists' Conference on the EPA Modeling Guideline," pp. 111-120, 22-24 February 1977, Argonne National Laboratory, Argonne, Illinois.

Roth, P. M., et al. (1976), "An Evaluation of Methodologies for Assessing the Impact of Oxidant Control Strategies," EF76-112R, Systems Applications, Incorporated, San Rafael, California.